High-throughput detection, annotation, and interpretation of genomic variants using next-generation sequencing data

by

Dewey J. Kim

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

Oct, 2015

Abstract

Recent advances in sequencing technology have made it feasible to use next-generation sequencing (NGS) to characterize the genomic landscape of an individual. Unlike microarrays, which are usually only capable of detecting variants that have been previously discovered in a population, NGS is capable of discerning both common and rare *de novo* variants.

Sequencing studies that involve the analysis of rare variants in human disease typically follow three steps: variant calling, where variants in NGS data are identified, variant annotation, where biologically relevant features are attached to each variant, and variant interpretation, where statistical and machine learning methods are used to prioritize putative functional variants. In this thesis, I attempt to apply and improve these methods in the context of cancer and schizophrenia.

Recent matched tumor/normal whole exome sequencing studies, coupled with current variant calling tools, have generated large sets of high-confidence genomic variants. A significant proportion of these variants are missense variants of unknown impact. To increase the speed and efficiency of annotating these variants,

ABSTRACT

I helped in the creation of a database of 86 precomputed disease-relevant features for all possible missense variants in the human exome. This tool allows for near-instantaneous annotation of any variant dataset.

A common limitation of using variant effect prediction software is the limited ability to infer the actual functional impact of a putative mutation. Most bioinformatics tools will only return a value signifiying the likelihood that the variant is functional. In an effort to aid in the interpretation of candidate mutations, I created a tool capable of mapping one-dimensional sequence positions and features onto three-dimensional protein positions. Using this database, coupled with an online web interface, an interactive 3D visualization of the variant position in the context of its protein is made possible.

I also utilized the NGS variant analysis pipeline to try and uncover novel insights into the genetics of schizophrenia using a matched case/control dataset of 500 postmortem human brain RNA-seq samples. Since RNA-seq is a relatively new NGS technique, I participated in a study to evaluate its technical reproducibility. As part of this study, I found that a new RNA-seq library preparation protocol, involving the depletion of ribosomal RNA using magnetic beads, allows for consistently high detection of intronic reads from pre-mRNAs and of long noncoding RNAs (lncRNA). To evaulate the role of rare non-coding variants in schizophrenia, I developed a pipeline for calling rare variants in RNA-seq data, involving the use of a series of alignment tools, resulting in an 80% decrease in

ABSTRACT

the number of false positives as compared to a standard approach. I also created a pipeline for detecting and analyzing short tandem repeats in RNA-seq data. Using this pipeline, I discovered statistically significant alterations in intronic short tandem repeats within genes involved in the innate immune response.

Primary Reader: Daniel Weinberger

Secondary Reader: Carlo Colantuoni

Acknowledgments

A large number of people need to be acknowledged in the creation of this dissertation. Instead of writing the names out in paragraph form, I created a table, which should render on the following page, in true LaTeX form.

ACKNOWLEDGMENTS

Person	Reason for acknowledgement
My family	Source of constant emotional support
Daniel Weinberger	Taking time out of his busy schedule to help a grad student think more like a scientist
Jooheon Shin	Helping me follow the yellow brick road
Jie Wu	That eQTLs are difficult
Hunki Lim	Sequence alignment is a hard, yet indispensable part of NGS
Rachel Karchin	Opening the door to the world of bioinformatics
Gary Gao	Helping me understand the nature of next-gen sequencing
Hannah Carter	That there is more to life than writing makefiles
Yunching Chen	Playing tennis with a bunch of amateurs
Andy Wong	Showing me that Wing Chun gives you superpowers
Violeta Beleva Guthrie	Keeping a critical eye on network models
Tri Ngo	To learn to stop worrying and love the PhD
Ron Shmueli	An indispensable ping pong partner
Iwen Wu	Helping me look at the job world more rationally
Ashwan Reddy	Making me want to move to DC
Amanda Price	Like a slow cooker, keeping our hearts & stomachs warm
Nina Rajpurohit	Showing us that we can bake ourselves out of any situation
Carrie Wright	That miRNAs are awesome
Genevieve Stein-O'Brien	That non-negative matrix factorization is awesome
Taeyoung Hwang	That RNA-editing is real
Bill Ulrich	That I'm not the only one who thinks Ubuntu is amazing
Zach Schultz	For convincing me to turn my acknowledgements into a table
Bert Vogelstein	For letting a hapless passenger learn to take the driver's wheel
Tychele Turner	That genomics and clustering seem to go together
Don Geman	Showing me the magic of random forests
Mark Diekhans	Showing me how bad I am at Python
Tamas Budavari	Showing me how bad I am at data science
Mike Ryan	Showing me how bad I am at SQL
Greg Hager	Showing me how bad I am at image analysis, and for tolerating me as a rotation student
Jonathan Pevsner	Being a part of my thesis committee, and help develop a critical eye to research
Kevin Yarema	Being a part of my thesis committee, and help look at the PhD in perspective
Carlo Colantuoni	Being a part of my thesis committee, and to consider NGS data in the time domain
Rebecca Birnbaum	That short tandem repeats are funfunfun
Richard Straub	Teaching me how to talk to scientists
Luca Ursini	Showing me how to be less of an idiot at the bench
Jean DuBose	For accommodating my unreasonable scheduling requests
Nicky Barat	For making sure I don't maim myself in the lab
Wan Hua Tan	For everything

Contents

List of Tables

List of Figures

Chapter 1

Using next-generation sequencing to probe human genetic variation in disease

The 21st century is proving to be an exciting time for bioinformatics: Since 2001, the amount of genomic data continues to rise exponentially, doubling every seven months — outpacing even Moore's Law [1]. This explosion of data, coupled with the continuously falling cost of computing, is paving the way for the development of new methods for detecting and analysing genetic (and genomic) signatures, which in turn allows for the generation of novel insights into human health and disease.

Disease is an unfortunate aspect of the human condition, and can be classified

into three parts: Infectious, environmental, and non-communicable. Infectious disease arises as a result of a transmissable pathogen; environmental diseases are a product of exposure to a toxin or poison. Non-communicable disease (NCD) is the global leading cause of death, responsible for 68% of all deaths in 2012 [2]. Almost all non-communicable diseases have a genetic aspect that either directly leads to illness or modulates an individual's predisposition towards illness. Identifying the underlying genetic variations is therefore a key component in uncovering the etiology of NCDs.

Sequencing is a powerful tool for identifying the primary structure (the nucleotides) of DNA or RNA sequences. Due to recent advances in sequencing technology, it is now possible to obtain a large subset of all genomic variants in a given genome, using suitable bioinformatics methods.[1]

This thesis will go over some of the efforts in the development of methods for analyzing genetic variants in next-generation sequencing data from cancer and schizophrenia. Chapter 1 will briefly go over the different types of genetic variation, next-generation sequencing, and the challenges of integrating data science with bioinformatics. Chapter 2 will discuss some of the basics of calling rare variants in NGS data, analysis of the feasibility of using RNA-seq data (including assessments on the reproducibility of RNA-seq), and will go over a bioinformatics

[1]Some types of variants are easier to detect than others, and the type of sequencing performed will also influence the kinds of variants that can be detected. This will be discussed in more detail in a later chapter.

pipeline I created in an attempt to call rare variants in RNA-seq data. Chapter 3 will provide a short overview of the genetics of cancer, and some heuristics I generated in an effort to characterize genes in cancer and human disease. Chapter 4 is adapted from a manuscript I co-wrote describing a database of pre-computed cancer-related features in the human exome. Chapter 5 is adapted from another manuscript I co-wrote of a database for mapping positions and features from 1D sequence space onto 3D protein structures. Chapter 6 provides an overview of the genetics of schizophrenia, and chronicles my attempts to discover genes that may be enriched for rare variants in schizophrenia. This chapter also goes over some of the quality control efforts I made when analyzing the RNA-seq data. In chapter 7, I explore the possible effects of short tandem repeats in schizophrenia, trying to see if there are statistically significant alterations in short tandem repeats in a schizophrenic population. I also delineate a computational pipeline for calling short tandem repeats in RNA-seq data. Chapter 8 discusses potential future work and ideas.

1.1 Human genetic variation

Humans, like all organisms, possess a large spectrum of genetic differences. Due to the stochastic nature of DNA duplication, even monozygotic ("identical") twins are genetically distinct from each other [3]. In this section, I will go through

a brief overview of the human genome and outline the different mechanisms of genetic variation.

1.1.1 The human genome

The human genome consists of 22 pairs of autosomes[2], and a pair of sex chromosomes. The autosomes are numbered from 1 to 22, in rough order of length, and the sex chromosomes are referred to as "X" and "Y". In total, the haploid human genome consists of around 3.2 billion base pairs [4].

The human exome is the most extensively studied subset of the human genome, consisting of around $20,000$ protein-coding genes and comprising about 1% of the human genome [5]. Due to alternative splicing, each gene can encode for several alternative transcripts. To date, there are approximately 80,000 identified protein-coding transcripts [6].

The existence of non-coding genes has been established for decades, but the prevalence of non-coding RNAs (ncRNA) in the human genome has only been recently established through the study of the transcriptome [7–9]. To date, there are approximately 25,000 non-coding genes, but this number is expected to change, as the functionality of the majority of these putative genes have yet to be determined [10].

[2]Autosomes are chromosomes that are not allosomes, or sex chromosomes

1.1.2 Types of variation

The human genome can vary through a variety of mechanisms, ranging from a single nucleotide change to gross chromosomal rearrangements or deletions. In this subsection, we'll go through the different types of variants.

1.1.2.1 Chromosome abnormality

Chromosomal abnormalities are large changes in a section of a chromosome's DNA. Aneuploidy is the most dramatic type of chromosomal abnormality, where either an organism is missing an entire chromosome or has an abnormal number of duplicate chromosomes. In a diploid organism (such as ourselves), *monoploidy* refers to a deletion event (where there is only one copy of a chromosome where there should be two), and *trisomy* refers to a single duplication event (where there are three copies of a chromosome[3]. Most cases of aneuploidy do not result in viable offspring, and result from improper cell division during meiosis. The most common disorder associated with aneuploidy is Down syndrome, which can be caused by having three copies (trisomy) of chromosome 21 [11].

Besides duplication or deletion of an entire chromosome, sections of a chromosome can also be deleted or duplicated. Translocations can also occur, where sections of DNA swap between chromosomes. In order to be classified as a chromosomal abnormality, the lengths of these alterations typically need to be at least

[3]Tetrasomy would refer to four copies.

a few megabases (Mb) in length, otherwise they are sometimes classified as structural variations.

1.1.2.2 Structural variation

Structural variations are typically defined as any change in the DNA sequence of length between one kilobase (Kb, or one thousand bases) to a few megabases. The most common type of structural variation is copy number variation (CNV), which involves duplications, deletions, insertions, or translocations of a swath of DNA. CNVs occur frequently in the human genome, and account for the most variability in the genome, when counting by base [12].

1.1.2.3 Insertions, deletions, and indels

The terms "insertion" and "deletion" in genetics typically refers to an insertion or deletion of DNA sequence of length 1 to 1000. The term "indel" is a bit ambiguous, and can either mean "insertion *or* deletion" or a specific type of variation where a swath of DNA sequence is first *deleted* and is then *replaced* by (via an insertion) a different DNA sequence. According to the 1,000 Genomes study, an individual has on average 344,000 insertions & deletions [13]. Due to technical limitations due to sequencing, this number may be an underestimate [14].

In the context of exome sequencing, with a focus on protein-coding regions of the genome, insertions and deletions tend to be differentiated into two subtypes:

frameshift and non-frameshift variants. Because translation involves codons of three nucleotides, an insertion or deletion that does not have a length that is a multiple of three will result in a shift in the subsequent coding alphabet, leading to a dramatic change in the amino acid output (and hence, a frameshift). Such changes do not occur when the length change is a multiple of 3, which will merely remove or insert amino acids.

1.1.2.4 Single nucleotide variant

Single nucleotide variants (SNVs) are the most prevalent type of variation in the human genome. SNVs can be divided into separate categories, depending on their minor allele frequency (MAF). SNVs that occur in more than 1% of the population tend to be considered common, and are often categorized as single nucleotide polymorphisms (SNPs). An average individual has about 3.6 million SNPs in their genome [13]. The number of "rare" SNVs (of minor allele frequency less than 1%) is still yet to be confidently determined, but I estimate this value to be no more than 300,000 per individual genome, based on original research.

Most variant prioritization tools have been focused on SNVs, and for good reason. SNVs are relatively easier to characterize than indels, and often times are easier to detect compared to structural variants.

1.1.2.5 Variable number tandem repeat

A variable number tandem repeat (VNTR) is composed of a DNA motif that is repeated a number of times in the genome. VNTRs are usually split into two categories, microsatellites and minisatellites. Microsatellites, also known as short tandem repeats (STRs), are comprised of motifs of lengths 1-6 base pairs, repeated anywhere between 5-50 times. Minisatellites have larger motifs, usually between 6-100 base pairs. There are approximately 700,000 STRs in the human genome [15].

1.2 Next-generation sequencing

Thanks to recent advances in sequencing, it is now possible to get a comprehensive readout of many types of genetic variations in a given genome. In order to understand some of the principles of detecting variants in next-gen sequencing (NGS) data, an understanding of the principles behind NGS is required. This section will provide a brief primer of NGS, and of the different types of NGS methods available.

The fundamental principle behind next-gen sequencing (NGS) is the same as shotgun sequencing: high-throughput parallel sequencing of random, short segments of the genome of interest. The main innovation driving NGS involves the "high-throughput" part — figuring out the DNA sequence of a massive number

of small fragments at the same time (in parallel). Different NGS technologies achieve this step through different means.

For Illumina sequencers, high-throughput sequencing is achieved through a method called "sequencing by synthesis." Sequence fragments are attached to the surface of a microfluidic "flow cell," and fluorescently-labeled nucleotides are allowed to pair themselves to the sequences in a stepwise fashion. During each step, replication of the sequence is advanced by one nucleotide. Each nucleotide is ligated with a differently-colored fluorophore[4], and a camera takes a picture each cycle. A flowcell can contain millions of sequences, which then appear as a constellation of multicolored "dots" on the camera. Since the camera cannot possibly detect the light coming from just a single sequence, local amplification of each sequence "island" is achieved through bridge amplification. In this way, the flowcell has "colonies" of sequences, where each "colony" (or "polony") is composed of thousands of identical sequences.

For PacBio sequencers, a nanophotonic chip is used, containing tens of thousands of tiny wells. Each well acts as a zero-mode waveguide, which essentially focus light into a very small volume. A polymerase enzyme with DNA is affixed to the bottom of the well, and during synthesis, bases with fluorophores attached to the phosphate (which allows for the fluorophore to naturally detach from the DNA), and the resulting light is picked up by the camera in real-time [16].

[4]Actually, this is no longer the case, since Illumina has now transitioned to two-color chemistry.

1.3 Genetics of Disease

Genetic disorders are categorized based on the number of genes that contribute to the disorder. The most straightforward-to-characterize disorders are single gene disorders, where mutations in a single gene are a necessary and sufficient cause to disease. Most mendelian diseases are single gene diseases, and fall under the following categories:

- *Autosomal dominant/recessive.* These genetic disorders involve genes in the autosomes (not chrM, chrX, or chrY). Autosomal recessive diseases require both genes (in both autosomes) to be mutated, while autosomal dominant diseases only require a mutation in one gene for the disease to occur.

- *X-linked disorders.* These disorders involve mutations in the X chromosome. Since males have only one copy of chrX, and females have two copies, the inheritance pattern is slightly more complicated than most autosomal dominant/recessive disorders. Most X-linked disorders follow a recessive inheritance pattern, meaning that these disorders tend to be much more prevalent in males than females (since females would have to carry mutations in both X chromosomes, while males only need a mutation in the one X chromosome).

- *Y-linked disorders.* These disorders are exceedingly rare, mainly due to the fact that chrY is an extremely small chromosome, coding for only a handful

of genes. Y-linked disorders can only be passed from father to son.

- *Mitochondrial disorders.* These disorders are based on mutations of genes in the mitochondrial genome (often labeled as "chrM" or "chrMT"). The mitochondrial genome is a circular genome of 16,569 base pairs [17]. Diseases involving the mitochondrial genome are passed down from mother to child, since the mitochondria themselves are conferred to the child via the mother's ovum.

Disesases that arise from a combination of mutated genes are commonly referred to as "complex," or "polygenic" diseases. Most common diseases, such as heart disease, diabetes, cancer, schizophrenia, bipolar disorder, and major depressive disorders are complex diseases. Since these disorders depend on a variety of genes, the inheritance pattern is rarely straightforward to determine.

A number of paradigms have been proposed in an attempt to characterize the genetic etiology of common disease [18]. The most famous of these paradigms is the common disease-common variant (CD-CV) hypothesis. The CD-CV hypothesis posits that the foundation of common disease lies in the combination of commonly-occurring genomic variants. The CD-CV hypothesis is what is driving GWAS (genome-wide association study), where millions of single nucleotide polymorphisms (SNPs) are polled for each individual. Although GWAS has been able to uncover many SNPs and genes that contribute to disease, they only seem to explain a small fraction of the inheritance of common disease — this is called

the "missing heritability" problem. Several ideas have been put forward to try and address the missing heritability: Some heritability might be overlooked by not considering interactions of common variants, and rare variants with high penetrance will also be overlooked when using GWAS [19,20]. Analyses of familial studies have often generated rare SNVs and structural variants, which led to the formulation of the common disease-rare variant (CD-RV) hypothesis — that a small number of highly-penetrant rare variants affect gene function, leading to disease.

1.4 Data science and bioinformatics

Traditionally, bioinformatics methods have been limited by the "small n, large p" problem, where n is the number of samples, and p is the number of "predictors," or features in a given sample. A typical example is a microarray study carried out on a limited number of subjects. Sample size may be extremely small ($n = 30$ subjects), but a microarray would generate a large p for each sample ($p = 20,000$ gene expression values). This problem has also been observed in other fields, such as image analysis, economics, and astronomy [21–23].

The meteoric rise in the amount of genomic information is a product of recent substantial improvements in sequencing technology. From 2007 to 2012, the cost of sequencing has fallen by over three orders of magnitude [24]. Due to the

dramatically lower cost of sequencing, it is now possible to generate "medium n, large p" datasets, where n is now approaching the same magnitude as p. This is great news when performing statistical analyses: Increasing the sample size will improve statistical power, but having datasets of such a magnitude presents challenges for data storage and organization.

For example, consider a typical case/control study utilizing transcriptomic data. There are currently 198,619 known, annotated transcripts [10]. If we have 1,000 cases and 1,000 controls, our study will generate 397,238,000 records. If we store this information into unsorted flat files (text files), it will take forever to try and search for necessary information.

Fortunately, bioinformatics data is often represented in a tabular format, which makes it easy to incorporate into a relational database. The dataset will be stored in a SQL (Structured Query Language) table; singular data records will occupy a row of the table, and features (or attibutes or fields) will be stored in the columns. For this example, the table would most likely have columns such as `sample_id`, `transcript_id`, `expression_value`, and each row would be a single expression value of a single transcript in a single sample.

Perhaps the most useful aspect of a relational database is the ability to put constraints on the data. Constraints such as enforcing the expression values to be non-negative acts to verify the integrity of incoming data. Indexing the data allows for fast data access, and it is akin to a library creating a card-catalog —

unlike flat files, where search times follow $O(n)$ time, indexing allows queries to follow $O(\log(n))$ time.

SQL databases are also easily scalable. The underlying architecture can be identical for a 10-record database and a 10-billion record database, with no singificant hit on performance.

Chapter 2

Calling rare variants in RNA-seq data

RNA-seq is traditionally used to quantify the expression of genes and transcripts, identify gene fusion events, and detect instances of RNA editing [25–29]. Since RNA-seq is still fundamentally a sequencing method, it is possible to analyze the read-out of the sequence and look for variants. In this chapter, I will evaluate the technical reproducibility of RNA-seq, identify potential sources of bias, go over the typical bioinformatics pipeline for calling rare variants, determine the fitness of the typical rare-variant calling pipeline on RNA-seq data, and propose a new method for calling rare variants in RNA-seq data.

2.1 Evaluation of RNA-seq

Because RNA-seq is a relatively new sequencing technology, it is vital to evaluate its technical reproducibility, as well as to identify possible sources of biases. In order to achieve this objective, we took a standardized sample of RNA, distributed identical copies of this sample to different sequencing centers, and asked various groups to sequence multiple technical replicates of the RNA using different sequencing platforms and RNA preparation protocols (Figure 2.1). This section is based on material from Li *et al.* 2014 [30].

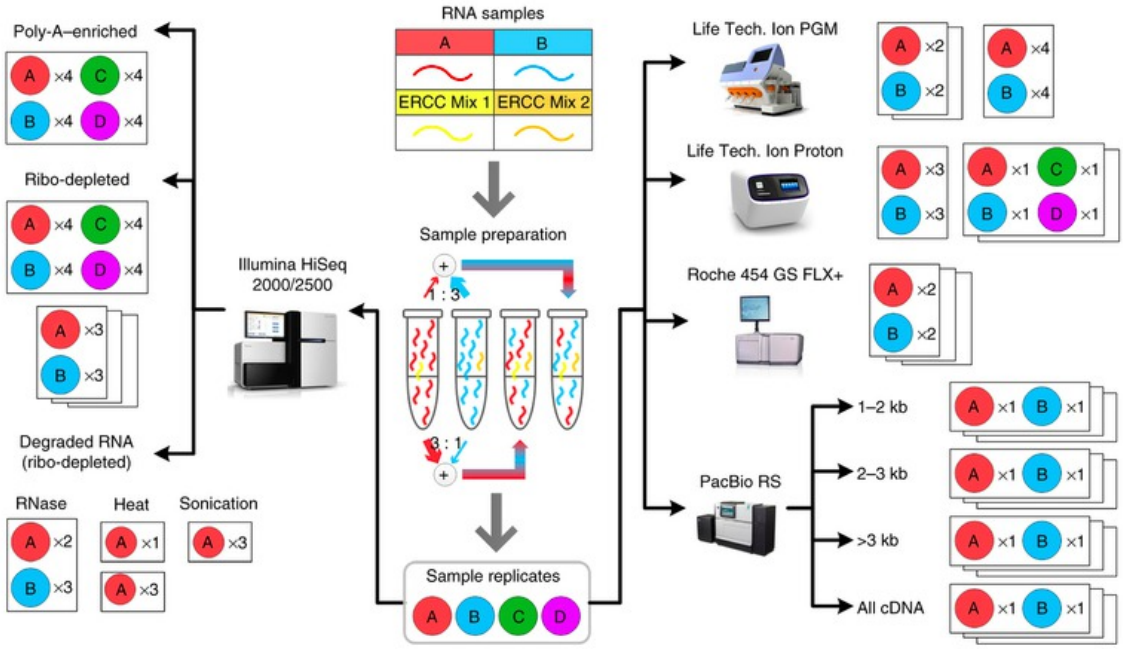

Figure 2.1: ABRF NGS RNA-seq study design.

2.1.1 Experimental design

Two separate commercial sets of standardized RNA samples were used in the study. The first sample, the Universal Human Reference RNA (UHRR), consists of a mix of total RNA from 10 human cell lines. The second sample, the Human Brain Reference RNA (HBRR), consists of total RNA from 23 Caucasians, taken from various regions of the brain. For this study, the UHRR sample was labeled as sample "A", and the HBRR sample was labeled as sample "B." These samples were previously used in a similar study evaluating the reproducibility of microarrays [31]. In addition to the A and B samples, 1:3 and 3:1 mixes of A and B were also created (labeled as "C" and "D").

Synthetic RNA of known primary structure and concentrations, developed by the External RNA Control Consortium, were also "spiked-into" each mix. These "spike-ins" consist of 92 polyadenylated transcripts, ranging from 250-2000 nucleotides, with a primary sequence designed to be distinct from the human genome (thus maximizing the chance that a read originating from the ERCC can be uniquely aligned) [32,33].

Technical replicates of each A, B, A/B mix were distributed to several sequencing centers, and were sequenced on a variety of platforms: Illumina HiSeq 2000/2500, Life Technologies Ion PGM, Roche 454, and Pacific Biosciences (PacBio) RS. Of particular interest are the reads originating from the Illumina HiSeq 2000/2500, since the data involved in this text are generated from Illumina machines.

2.1.1.1 RNA preparation

In order to prepare the RNA for sequencing on the HiSeq 2000/2500, two different library preparation protocols were followed. The first protocol used the Illumina TruSeq Library Preparation Kit, which selects for polyadenylated transripts. The second protocol used the Illumina TruSeq Library Preparation Kit, paired with the Epicentre Biolotechnologies Ribo-Zero Gold system, which removes ribosomal and mitochondrial RNA.

2.1.1.2 Sequencing

After the DNA was confirmed to be of sufficient quality by Bioanalyzer analysis, the completed libraries were attached with barcodes, multiplexed to 12 samples per lane, and distributed across channels on three different flow cells, in order to account for variability due to lane number and run sequence. Paired-end 2x50bp reads were generated for each run, the raw BCL (base call) files were then demultiplexed, converted to fastq, and then distributed to the members of the consortium.

2.1.2 Bioinformatics

Quality control (QC) information was first generated on the fastq reads using fastQC [34]. The raw fastq reads were then aligned to the hg19 (GRCh37) hu-

man assembly using the STAR aligner [35]. With the resulting BAM files, gene expression values (FPKMs) were generated using htseq-count [36].

2.1.2.1 RNA-seq quantification metrics

For subsequent downstream analysis of gene quantification, FPKM values were used, which stand for "Fragments per kilobase of exon per million mapped reads." In this segment we will define this, and other similar quantification metrics.

The RPKM/FPKM metric (reads/fragments per kilobase of exon per million mapped reads) was designed to account for a length bias in the number of reads that map to a transcript/gene of interest. Given m genes/transcripts, the RPKM metric for a given gene/transcript i is defined as

$$\text{RPKM}_i = \frac{C_i}{\left(\frac{\tilde{L}_i}{10^3}\right)\left(\frac{\sum_{j=1}^{m} C_j}{10^6}\right)} \tag{2.1}$$

where C is the number of reads (counts), and $\tilde{L}$ is the effective length of the gene/transcript of interest, which is defined as

$$\tilde{L} = L - \mu_L + 1 \tag{2.2}$$

where L is the length of the gene/transcript of interest, and μ_L is the mean read/fragment length. The effective length is used since the number of possi-

ble positions a read can map to a gene/transcript depends on the length of the read, as well as the length of the transcript.

For example, take the string "ABCDEF." If we are sequencing reads of length 3 from this string, there are $6 - 3 + 1 = 4$ possible reads that map to "ABCDEF": "ABC," "BCD," "CDE," and "DEF."

Equation 2.1 can be simplified into the following:

$$\text{RPKM}_i = \frac{C_i}{\tilde{L}_i \sum_{j=1}^{m} C_j}\,. \tag{2.3}$$

For paired-end reads, a similar metric, the FPKM (fragments per kilobase of exon per million mapped reads) is used, which treats each read of a pair separately. In short, for single-end reads, the RPKM metric is used, and for paired-end reads, FPKM is the metric to use.

Due to some criticism of the RPKM/FPKM metric's inconsistency when comparing across samples, the transcripts per million (TPM) metric has gained some popularity [37,38]. The TPM metric is defined as:

$$\text{TPM}_i = \frac{C_i}{\tilde{L}_i} \left(\frac{1}{\sum_{j=1}^{m} \frac{C_j}{\tilde{L}_j}} \right) 10^6\,. \tag{2.4}$$

The TPM has an added benefit of being easily interpretable; TPM_i is the number of transcripts one would expect to see for transcript i when one randomly sequences 1 million transcripts from the sample.

2.1.2.2 Between-sample vs. Within-sample normalization

The above defined metrics, the RPKM/FPKM and TPM, are an attempt to normalize the expression values within a given sample. Unfortunately, these metrics tend to have difficulty comparing expression between distinct samples [39,40]. In order to successfully compare expression samples, between-sample normalization methods have been developed — the most popular methods are Trimmed Mean of M values (TMM), and the DESeq normalization package [41,42].

2.1.2.3 Differential expression analysis

In order to perform differential expression analysis on the technical replicates, raw counts were normalized using the Trimmed Mean of M values (TMM) method [42].

2.1.2.4 RNA QC collection

In order to calculate more RNA-seq QC metrics downstream from alignment, the RSeQC package was used to calculate metrics such as the genebody coverage, read distribution, duplication rate, and the chrM read proportion [43]. RSeQC takes in a sorted, indexed BAM file, and returns a folder of metrics, depending on the type of test desired.

2.1.2.5 Variability of RNA-seq transcript quantification

In order to assess the variability in transcript expression values compared across technical replicates and across sequencing platforms, the coefficient of variation was calculated for each normalized transcript count. The coefficient of variation is defined to be the ratio of the standard deviation against the mean:

$$c_v = \frac{\sigma}{\mu} \tag{2.5}$$

For empirical data, the coefficient of variation can be estimated using the sample standard deviation and the sample mean:

$$\hat{c_v} = \frac{s}{\bar{x}} \tag{2.6}$$

When comparing between platforms, the Illumina HiSeq generates the lowest median inter-site variation (Figure 2.2). This low degree of variation exhibited by Illumina data is also demonstrated when comparing the Spearman correlation (0.98 for all samples of "A", and 0.99 for all samples of "B"). Due to the extremely high number of reads generated, the HiSeq also demonstrates the highest level of sensitivity when detecting genes. Unfortunately, due to the extremely short read length (50bp in this study), the junction detection rate is lower than that of other platforms. A separate study has verified that increased read length significantly improves junction detection (leading to better assembly accuracy), but that single-

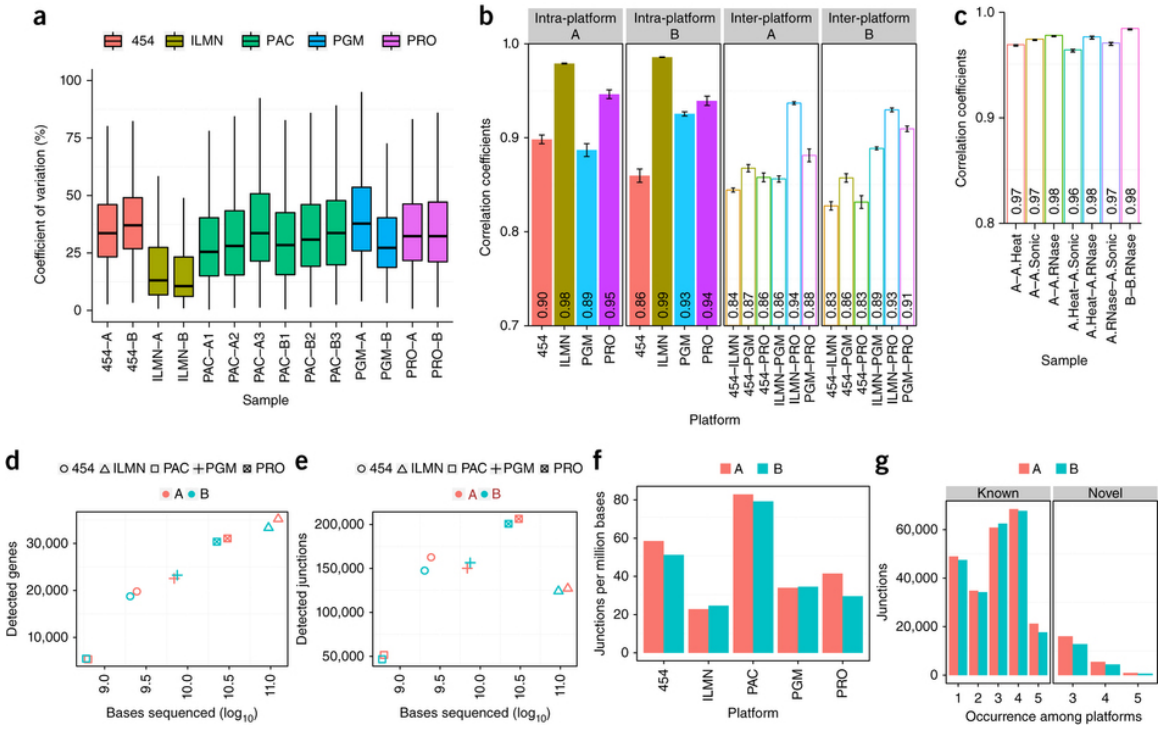

Figure 2.2: Variation metrics of normalized transcript quantification, evaluated within and between sequencing platforms. (**a**) Between-site coefficient of variation (CV) of normalized transcript expression values, evaluated across all sites. (**b**) Intra- and inter- platform Spearman correlation coefficients for samples A and B. (**c**) CV of samples matched against degraded RNA samples. (**d,e**) Number of detected genes & junctions against number of bases sequenced. (**f**) Average read length improves junction detection rate. Illumina has the shortest read length (50bp), and PacBio has the longest read length (several kilobases). (**g**) Already annotated junctions are picked up more reliably than "novel" junctions.

end 50bp reads are sufficient for performing differential expression analyses [44].

When comparing identical RNA samples enriched through Poly-A and RiboZero, there is a reasonably high degree of correlation of expression values, with Spearman correlations ranging between 0.91 and 0.93.

2.1.2.6 Read distribution

RSeQC was used to determine the percentage of reads mapping to various gene regions [43]. When comparing Illumina RNA samples that were prepared by poly-A pulldown vs. RiboZero, we noticed significantly more intronic reads coming from the RiboZero data (Figure 2.3). This is due to unspliced pre-mRNA reads originating from the nucleus [45,46]. There is also an uptick in the number of ribosomal RNA detected in the RiboZero data, whereas the PolyA data has virtually no ribosomal reads. This will be an important factor to consider for RiboZero data; filtering out rRNA reads will be necessary to have comparable expression values downstream, especially if the rRNA read proportion is greater than 1%.

2.2 Standard pipeline for calling rare variants in NGS data

The previous section went over the technical reproducibility of RNA-seq data, as well as some of the differences between Poly-A and RiboZero libraries. In this section, I will go over the typical bioinformatics tools used to ascertain rare variants from raw NGS data.

When calling variants from NGS data, whole genome sequencing (WGS) or

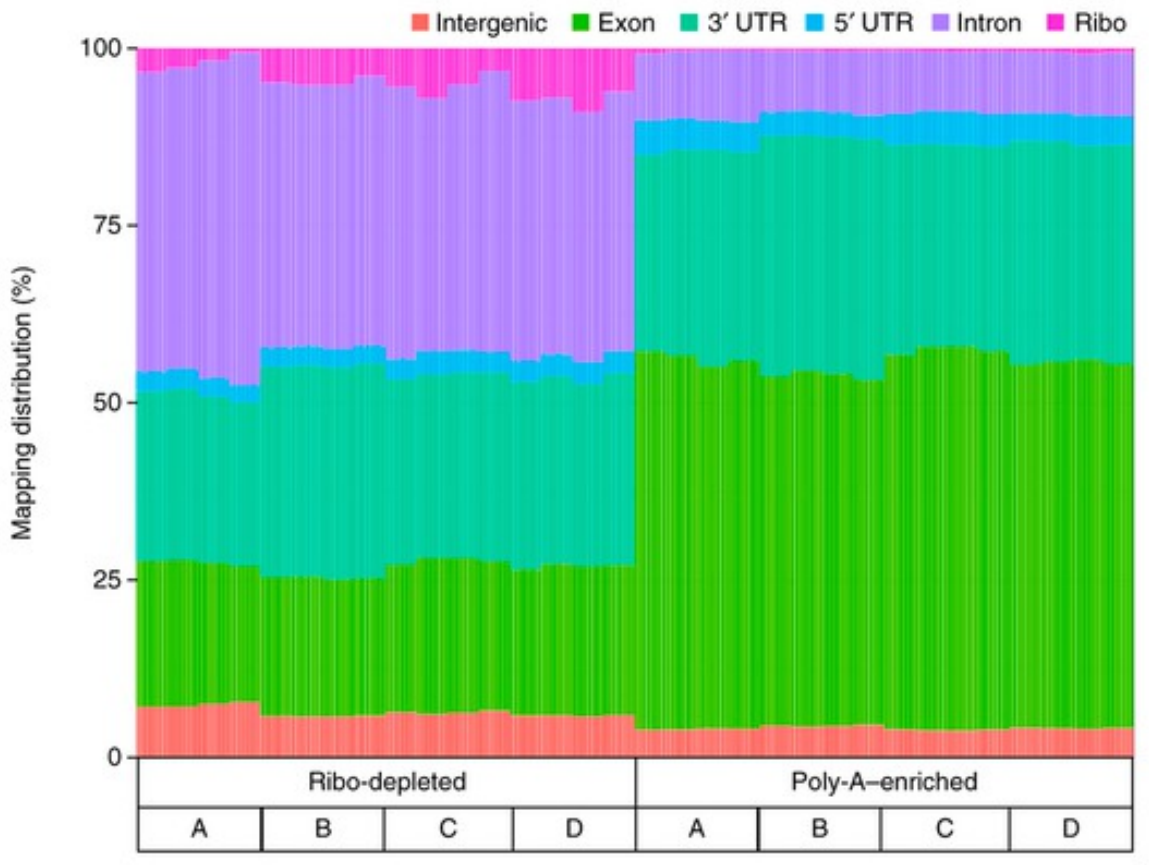

Figure 2.3: Distribution of Illumina reads mapping to different gene regions, from reads coming from Poly-A enriched samples and RiboZero depleted samples.

whole exome sequencing (WES) data is typically used. Due to the significant cost savings of WES (as compared to WGS), if variants in coding regions are of primary interest, then whole exome sequencing is the traditional tool to use. However, recent research has challenged this prevailing line of thought [47]. Coding SNVs seem to be well-detected using both methods, but CNVs remain a challenge for WES.

The type of technical considerations that need addressing will depend on the type of variant one is interested in calling:

1. *Structural variants.* The detection of structural variants (SVs) relies on the existence of three types of reads:

   (a) *Depth of coverage.* For WGS and WES data, CNVs can be detected by finding regions of unusually high or low coverage.

   (b) *Split reads.* A read that, when split, maps to two disjoint genomic regions, could be an indication of a SV.

   (c) *Discordant read pair.* A paired-end read, where each mate maps to vastly distant regions of the genome, can also be an indication of a SV.

2. *Small insertions/deletions, and SNVs.* To check the veracity of a candidate insertion/deletion or SNVs, there must be sufficient coverage of the region of interest to make sure that the putative variant is not a false positive that may have arisen due to a single read's sequencing error, or perhaps due to mistakes during PCR (polymerase chain reaction).

3. *Variable number tandem repeats.* In order to determine the full length of a VNTR, the sequencer must be able to generate reads that can fully span the length of the repeat, as well as its flanking region. The length of the sequenced flanking region(s) must be sufficiently long enough to be uniquely mappable to the reference genome. Technical errors must also be taken into account, as the rate of variation of a VNTR is many orders of magnitude higher than that of any other variant type [48–50]. For paired-end reads, it

may be possible to determine the STR length if the STR happens to span both mates, and the length of the fragment is known.

In order to limit the scope of this thesis, we will shift our focus towards calling STRs, short insertions/deletions, and SNVs. But before I completely abandon SVs, a few remarks about calling structural variants using RNA-seq:

1. CNVs will be extremely difficult to detect in RNA-seq data, as these may be indistinguishable from simple over/under-expression.

2. Gene fusion events can definitely be detected using RNA-seq, and RNA-seq may possibly be an ideal tool for detecting fusions.

3. There are aligners that are tailored for RNA-seq reads, that can also detect chimeric reads (reads mapping to separate places in the genome). TopHat & STAR are two such aligners that come to mind.

2.2.1 Bioinformatics pipeline for calling rare variants

This subsection will go through the typical software used to process rare variants from raw NGS data. In short, the pipeline is as follows: QC, alignment, more QC, variant calling, variant filtering, variant annotation, variant interpretation/prioritization.

2.2.1.1 Raw fastq quality control

Before performing variant calling (or any sort of bioinformatics analysis), we need to make sure that the raw fastq files coming out of the sequencing machines meet certain quality standards. FastQC is the tool of choice, which is a java program that runs a battery of QC checks on the fastq file [34]. The following is a list of commonly-encountered considerations when dealing with non-ideal fastqc files:

- *Adapter sequences.* Adapter sequences are usually trimmed prior to preparation of the fastq file. If this step was not performed correctly, adapter sequences in the fastq files may be misconstrued as false insertions or false SNVs.

- *Low base quality.* Each nucleotide in a fastq file has a corresponding quality score, which reports the likelihood of the particular base being called incorrectly [51]. This "phred" score Q is defined as $Q = -10 \log_{10} P$, where P is the probability of an incorrect base call. A phred score of $Q = 40$ would then mean an error probability of $P = 10^{-40/10} = 10^{-4} = 1/10000$. Low phred scores will result in false positive variant calls. For Illumina sequencing data, the base quality scores often take a dive near the end of the sequence, so care must be taken when calling variants from the ends of a read. If the quality scores near the read ends are too low ($Q < 25$ is a rule of thumb), it

may be advisable to trim the ends of the reads.

- *Duplication rate.* For RNA-seq data, it is natural to have some duplicated reads due to highly expressed genes. If the number of duplicated reads is inordinately high, say, greater than 10%, then this could be a sign of a serious PCR amplification artifact.

2.2.1.2 Read alignment

Once the fastq files have been confirmed to be of sufficient quality, the next step in the bioinformatics pipeline is to align the reads to a reference genome. In other words, we need to know where in the genome a read maps to. For human fastq files, the typical reference is either hg19 (GRCh37) or hg38 (GRCh38). These represent the "standardized" haploid sequences of all human chromosomes. Hg19/hg38 does not represent an actual human individual, it is an attempt to "average out" the variation of a diverse population of human genomes. This is, of course, an imperfect process, so what is considered a "variant" when called against hg19/hg38 may actually be the most prevalent allele in the population.

For WGS and WES, there are several aligners that are capable of mapping reads to a reference in a very short amount of time, compared to traditional aligners such as Smith-Waterman, Needleman-Wunsch, BLAST, and BLAT. Bowtie and BWA are the two most commonly-used aligners to date.

For RNA-seq reads, performing high-quality alignments is significantly more difficult than aligning short reads from WGS and WES data, due to the fact that there will be reads that span exon-intron junctions. To an unprepared aligner, these will look like giant deletions, and indeed, most WES/WGS aligners make a key assumption that most reads will be contiguously mappable to the reference genome. TopHat is a popular RNA-seq aligner, but STAR is beginning to become quite prominent as well, due to its significantly faster rate of alignment [35,52].

2.2.1.3 BAM file QC

Once fastq reads have been mapped to a human reference, the typical output is a Sequence Alignment Map (SAM) file. In practice, the BAM file, a compressed version of the SAM file, is stored after alignment.

Once the BAM file is generated, we can run RSeQC, *samtools flagstat*, and the aligner QC reports, to assess other quality control metrics:

- *Mapping rate.* If an inordinate proportion of reads are not mappable to the reference genome (below 80% is a rule of thumb), this is a red flag, and could indicate the following problems:

    1. RNA/DNA contamination from another species
    2. Enrichment of reads of extremely low quality, rendering them unmappable

3. Reads dominated by adapter sequences
4. Reads originating from regions of the reference genome that are not uniquely mappable

- *Proportion of reads originating from chrM.* An overabundance of reads originating from the mitochondrial genome may indicate severe degradation of the sample of interest, or a failure of library construction.

- *Proportion of rRNA reads.* Unique to RNA-seq, an overabundance of ribosomal reads may also indicate a failure of library construction, or a sign of a severely degraded sample, where only rRNA is present.

2.2.1.4 BAM file processing

Once the BAM file has been sorted and indexed, which is typically done with *samtools*, additional processing needs to take place, depending on the type of sequencing. For exome and WGS data, in order to minimize the impact of PCR amplification errors, duplicate reads need to be removed from the BAM file. *samtools rmdup* or *Picard* are tools that are capable of flagging and removing duplicate reads.

For calling short insertions, deletions, and indels, it may also be preferable to re-align reads with putative insertions/deletions/indels, as some of these candidate variants may be simply false positives arising from misalignment. Local

alignment is also preferable when one is interested in SNVs that occur near indels, since a falsely called indel can also result in false positive SNVs being called nearby. The *IndelRealigner* tool from the Genome Analysis Toolkit (GATK) is one such tool for realigning indels.

2.2.1.5 Variant calling

With the BAM file properly processed, we can use variant calling software to get a list of putative variants [53]. For exome and WGS data, the most popular tools are as follows:

- *Samtools mpileup*. The first variant caller that gained widespread use, *mpileup* is relatively fast and efficient, but the false positive filters are a bit primitive and the threshold for calling variants is quite optimistic, leading to great sensitivity, but lower specificity, especially when calling indels [54].

- *Genome Analysis Toolkit's UnifiedGenotyper & HaplotypeCaller.* Genome Analysis Toolkit (GATK) is currently considered the golden standard for variant calling. GATK's UnifiedGenotyper uses a Bayesian model, whereas the GATK HaplotypeCaller first re-assembles the candidate variant region to determine candidate haplotypes, re-aligns using Smith-Waterman, then determines the likelihood of a particular haplotype by performing a pairwise alignment of reads against each candidate haplotype. Genotypes are then

calculated by selecting the alleles with the highest likelihoods (given each haplotype). Currently, GATK recommends the use of HaplotypeCaller over UnifiedGenotyper [55].

- *Freebayes.* Freebayes is another haplotype-based variant calling tool, where haplotypes are estimated with a Bayesian approach. Freebayes is capable of phasing, modeling multiallelic loci, handling non-diploid organisms, and can detect complex multinucleotide variants (MNVs) [56].

- *Platypus.* Platypus is another variant caller that also calls variants using local assembly, realignment, and haplotype estimation using a Bayesian model. The main advantage of Platypus is the fact that it is at least 10 times faster than other variant callers [57].

2.2.1.6 Variant filtering

Once a candidate variant is called, various filtering steps must be followed in order to make sure the variant is not a false positive. Ideally, variant calling is an extremely easy task — simply find areas that mismatch against the reference. Unfortunately, the majority of such mismatches are due to false positives.

There are two approaches to dealing with false positives. One is to incorporate the false positive likelihoods into an already existing Bayesian model. Another is to use a battery of simple heuristic cut-offs that, in an ideal situation, can perfectly

separate false positives from true positives.

The following is a list of features that are taken into consideration when determining a variant's chance of being a false positive:

- *Base quality.* False positives tend to have lower base quality scores.
- *Unmapped mate.* A well-behaving paired-end read would ideally have both mates mapped to the same general location. If only one mate is mappable, that could be an indication of an improperly mapped read — that is, a read that shouldn't be mapped.
- *Strand bias.* A true, well-behaving variant should be invariant towards read strandedness; it should not matter which way the read goes, the variant should exist in both directions. A strong bias towards one strand could be a sign of an improper alignment.
- *Allele bias.* For a diploid organism, a variant should be either homozygous or heterozygous for an allele — that is, have exactly zero, one, or two copies of the alternative allele. Therefore, the proportion of reads containing alternative allele and the reference allele should ideally be 0%, 50%, or 100%.
- *Positional bias.* For a true variant, there should be no bias in the position of the detected variant inside the read. For example, if all the reads supporting a putative variant have the putative variant near the end of the read, this could be a sign of improper alignment — meaning that only reads of that

particular orientation are mapped to the putative region, resulting in a false positive call.

- *Mismatch bias.* It is unusual to encounter large clusters of variants in the same place (or in the same read), unless the read is improperly mapped, and generates several spurious mismatches.

These features are included in almost all variant calling tools, so manual variant filtering is not usually carried out. That being said, it is still important to be familiar with these features, since you can tweak the cutoffs to best suit your experiment, when running the variant calling tools. For example, if you have cancer data that has matched tumor/normal samples, and you are only interested in somatic mutations, having an initial high false positive rate might be acceptable, since you have the matched normal sample to nullify any false positives. On the other hand, if you are running a large case/control study, you might be willing to sacrifice sensitivity for specificity - in which case, you would run the caller with more stringent parameters.

2.2.1.7 Variant annotation

After running a variant caller, and getting a list of vetted, putative variants, these variants are usually stored in a Variant Call Format (VCF) file. A VCF file is a tab-delimited file that contains basic variant information, such as the chromo-

somal position, reference and alternative allele, and other optional information that is dependent on the variant caller used — such as depth of coverage, allele frequency, and quality score. [58,59].

Before performing the actual variant annotation, additional VCF processing steps might be necessary. Two popular tools for processing VCF files are vcftools, vcflib, and GATK [55,58,60]. The following is a list of common VCF parsing steps:

- *VCF filtering.* Depending on the type of VCF caller, a variant that fails filtering will have the failure mode flagged in the "FILTER" field, instead of being removed from the VCF file outright. For downstream annotation and analysis, it is often preferable to have these completely removed from the VCF.

- *Masking.* Masking areas of the genome, and "blacklisting," or removing all candidate variants that occur in these regions, is a common step in parsing VCF files. The most common type of mask is removing variants in very low entropy regions (such as areas annotated by repeatmasker [61].

- *Breaking or creating multiallelic variants.* Depending on the variant caller or the variant annotator used, it may be desirable to either re-create a previously broken multiallelic variant into a single line, or to create separate VCF records for multiallelic variants. It may also be desirable to only focus on the most common allele.

- *Breaking or creating multinucleotide variants.* Also dependent on the variant caller & annotator chosen, you may also wish to consider breaking up MNVs into adjacent SNVs, or joining adjacent SNVs into a single MNV record.

- *Left aligning indels & complex variants.* Most, but not all, variant callers will left-align variants. Since there are multiple ways to represent an indel, left-aligning an indel (changing its start and stop position to be closest to the 5′ end of a chromosome) is critical, since having multiple representations of the same variant will make downstream indel comparisons almost impossible [62].

- *VCF sorting and indexing.* If the VCF file is extremely large in size, some downstream tools will enjoy a performance boost if the VCF is sorted and indexed.

With a properly parsed VCF file, the next step is to annotate the variants, and attach a battery of features in order to get a better understanding of the variant in question. The most popular tool in use for variant annotation is *ANNOVAR*, but there are also other tools that are gaining popularity, such as *GEMINI* and *VAT* [63–65].

The following is a list of the most common annotations attached to variants:

- *Gene*. When doing downstream analyis, it is often critical to know which variants occur in which gene. If the variants are in intergenic regions, they

are often removed from further analysis.

- *Gene substructure.* If you are dealing with whole exome sequencing data, you are most likely interested only in variants in coding regions. If the variants are annotated to occur in untranslated (UTR) regions or introns, they may also be removed from further analysis.

- *Amino acid change.* When prioritizing variants in coding regions, it is often necessary to also know if there are any changes to the amino acid sequence. If there is no change (a synonymous variant), these variants may also be removed.

- *Conservation.* If a variant occurs in a highly conserved region, that might increase the chance of the variant negatively affecting downstream function.

- *Functional annotation.* Conservation is just one of a large variety of features that attempt to assess the functionality of a given genetic region. If one is interested in epigenetics, CpG islands may be of interest. If you are interested in alternative splicing, annotation of splicing motifs may be useful.

If calling rare variants is of primary interest, it is at this step where you would determine whether a called variant is known to be common, and thus would filter them out. Databases such as dbSNP, variants from 1,000 genomes, HapMap, are useful for identifying common variants.

2.2.1.8 Variant interpretation

Attached with an informative set of features, machine learning methods can be used to predict what, if any, impact the variant will have on protein or transcript function. Currently, most software packages are focused on predicting the impact of missense variants on protein function. The most well-known packages that do this are SIFT, Polyphen, SnpEff, and Variant Effect Predictor (VEP) [66–69]. There are also newer packages that are gaining in popularity: Variant Annotation, Analysis, and Search Tool (VAAST) and Functional Analysis through Hidden Markov Models (FATHMM) [68,70].

2.3 Unique strengths & challenges when using RNA-seq data to detect genomic variants

RNA-seq is primarily a tool for transcriptome characterization, but the raw data coming from the sequencer can still be used to call genomic variants. Calling variants using RNA-seq offers a couple of unique advantages over more traditional NGS methods:

1. *Coverage of non-coding regions.* Exome sequencing is designed to capture only

the coding region of the human genome, leaving out possibly interesting regions such as introns, UTRs, and previously unannotated non-coding RNAs. RNA-seq is capable of sequencing pre-mRNAs that have yet to have their introns fully spliced out.

2. *Cheaper than whole-genome sequencing.* Using high-coverage RNA-seq could be a decent compromise between cost and sensitivity when it comes to detecting non-coding genomic regions. With unlimited resources, whole genome sequencing (WGS) would be the ideal tool to use for calling non-coding variants, since it is unbiased in terms of the genomic region being sequenced. This absence of bias also is the reason behind its high cost, since many reads will be wasted sequencing uninteresting intergenic portions of the genome.

Before one goes about calling variants in RNA-seq, several challenges unique to RNA-seq data must be addressed:

1. *Non-uniform coverage.* RNA-seq is a functional read-out of the transcriptome. Therefore, genes that are not currently being transcribed will not be detectable in the RNA-seq data. Genes that are lowly expressed will also face difficulty in sequencing, therefore reducing the sensitivity of any downstream variant caller.

2. *Spliced reads.* RNA-seq will contain reads that span exons (and thus skip-

ping the intron). To an unprepared variant caller, this may look like a giant deletion.

3. *RNA editing.* RNA editing will be detectable in RNA-seq data, and to a sequencing machine, these events will be indistinguishable from SNVs.

In order to adapt standard rare-variant calling practices on RNA-seq data, I decided to make the following changes:

1. *Split spliced reads.* By convention, an RNA-seq aligner will flag spliced reads with an "N" in the CIGAR value, which is a string (in a SAM/BAM file) that summarizes the read alignment against the reference genome. Variant callers are unable to deal with "N" CIGAR values, and sometimes treat them as unacceptably long indels (and therefore discard these information-containing reads). To get around this issue, I used the tool *SplitNCigarReads* from GATK and *splitNReads* from Miika Ahdesmaki [55, 71].

2. *Remove RNA editing sites.* Once candidate variants are called and placed in a VCF file, I created a mask around known and predicted RNA-editing sites from the RADAR and DARNED databases [72, 73]. With this mask, I removed all putative rare variants that overlap with these positions.

In the following subsection, I will outline in detail the steps and packages I used to assemble the first prototype of the rare variant calling pipeline for RNA-seq.

2.3.1 Rare variant calling pipeline for RNA-seq, in detail

First, the raw fastq files were quality checked using fastQC. The quality-checked fastq files were then aligned to hg19 with GENCODE gene annotations (version 19). STAR aligner was used to align the reads. I chose STAR over other, possibly more accurate aligners, such as TopHat & GSNAP, due to STAR's overwhelming performance advantage: STAR is approximately 150 times faster than TopHat (and GSNAP is slower than TopHat).

Once the BAM files are generated, I sorted and indexed the files using *samtools*. To take care of the split reads, I used GATK's *SplitNCigarReads* to divide up junction-spanning reads. I then used GATK HaplotypeCaller to make the variant calls. On later runs, I also used freebayes and Platypus to call variants.

With the generated VCF files, I filtered out regions overlapping repeatmasker regions and RNA-editing sites predicted by RADAR and DARNED. The remaining variants were annotated by ANNOVAR, and common variants with a minor allele frequency $> 1\%$ occurring in 1000 genomes, the exome sequencing project (ESP6500), the exome aggregation consortium (ExAC), and Complete Genomics (CG46), were filtered out.

2.4 Evaluation of rare-variant calling pipeline on exemplary data

When inventing a new computational pipeline, it is advisable to evaluate the pipeline on a set of exemplary data, in order to measure valuable performance metrics such as the sensitivity and specificity. Here is a quick overview of some performance terms:

- *Sensitivity*, or the true positive rate, or recall, is the ratio of the number of samples identified as positive, divided by the total number of true positives.
- *Specificity*, or the true negative rate, is the ratio of the number of samples identified as negative, divided by the total number of true negatives.

In other words, sensitivity is a measure of how well a classifier can pick out true positives. Specificity is a measure of how well a classifier can avoid false positives.

I used the Genome in a Bottle (GIAB) consortium's extensively verified set of variants in sample NA12878 as the exemplary dataset for rare variant calling [74]. In the GIAB study, a human volunteer had her genome sequenced using extremely deep whole genome sequencing, using 5 different sequencers, 7 different aligners, and 3 different variant callers. A list of concordant genotypes was generated, with a corresponding BED file reporting confidently called regions.

In order to compare the RNA-seq rare variant calling pipeline with the GIAB dataset, I used publicly-available RNA-seq data of the same individual (NA12878), made available by the ENCODE consortium, sequenced on an older Illumina Genome Analyzer IIx (GAIIx), generating 2x76bp paired-end reads [75].

I called rare variants on this ENCODE RNA-seq dataset, and compared those against the exemplar variants detected by the GIAB consortium. The subsequent analysis generated surprising results — most of the putative rare variants called by a typical naïve variant calling pipeline generated a significant number of false positives. In fact, the number of false positives outnumbered the number of true positives. I observed similar results when calling variants using freebayes and Platypus. I also observed similar false positive rates when using a separate RNA-seq dataset (of NA12878).

The extremely high false positive rate is disturbing, but after more consideration (and some extensive literature review), seems plausible. Variant callers, when assessing performance, often combine common and rare variants together, making the assumption that the sensitivity and specificity of detecting common variants and rare variants are identical. Unfortunately, I believe that conflating the two variant types is dangerous when assessing rare variant calling performance. If a false positive call is made, what is the likelihood that it just happens to occur at a SNP? It will most likely be a unique false positive, and therefore be classified as a rare variant. A typical individual has approximately 3.5 million SNPs [13].

Individual NA12878, sequenced by GIAB, has 303,583 rare variants. Since the number of SNPs outnumber rare variants by an order of magnitude, conflating common and rare variants will generate an overly-optimistic estimation of the false positive rate, when calling rare variants.

2.5 Analysis of putative rare variants confirmed to be false positives

To gain a better understanding of the mechanisms behind the false positives, I used the Integrated Genomics Viewer (IGV) to open up the BAM files and visualize the reads that contribute to a given false positive [76]. IGV is a great way to visualize the distribution of reads mapping to a chromosomal region of interest. By default, any mismatches from the reference genome will be highlighted, low quality bases will be flagged (and greyed-out), and a histogram representing the depth of coverage at each base will be shown above the visualization of reads. I also mapped suspect reads against BLAT to determine the level of misalignment [77]. BLAT is capable of handling spilced reads, and has excellent sensitivity, and therefore can come up with a comprehensive list of alignment alternatives.

After manual inspection of over 500 false positives, I determined that most false positives arise from the following factors:

1. *Misalignment.* Re-alignment of a suspect read, using BLAT, reveals a better, mismatch-free alignment for the candidate read.

2. *Low base quality.* Consistently low base quality scores in the candidate region correlate with false positives.

3. *Soft clipped reads.* Aligners will often soft-clip reads with unsatisfactory alignment near the ends of a read. Soft-clipping is different from hard-clipping in that the sequence is preserved, but flagged (with a "S" in the CIGAR field). Sometimes, the aligner is not parsimonious enough with the soft-clipping, and permits low-quality regions to persist in the read.

4. *Homopolymer run.* A homopolymer run is a sequence of the same consecutive bases (for example, "AAAAAAAAAA"). Due to the extremely unstable nature of these runs, it is commonplace to observe expansions and contractions of these regions.

We can also see that these sources of error are also the underlying factors behind many of the features used to model false positives in variant calling tools. Table 2.1 gives a rundown of the underlying causes behind each false positive filter.

Table 2.1: False positive filters and their underlying causes.

False positive filter	Underlying cause
Base quality bias	Low base quality
Duplicated reads	Low base quality
Unmapped mate bias	Misalignment
Strand bias	Misalignment
Allele bias	Misalignment, RNA-editing
Positional bias	Misalignment
Mismatch bias	Misalignment, low base quality
Homopolymer region	Homopolymer expansion/contraction

2.6 Development of a new RNA-seq rare variant calling pipeline

With the drivers behind rare variant false positives in mind, I decided to come up with a new, more stringent rare variant calling pipeline that addresses each of the four main challenges.

1. *Remove soft-clipped reads.* Instead of trying to estimate how much of a soft-clipped read to trust, I simply eliminated all reads with a soft-clipped region from further analysis.

2. *Enact stricter base quality cutoffs.* In order to tackle false positives arising from lousy base quality, I tuned the variant caller to be more stringent when considering low quality bases.

3. *Mask homopolymer regions.* I created a BED (browser extensible data) contains all known homopolymer repeat regions in the genome. With this mask,

I removed all variants overlapping these regions.

4. *Re-align all reads containing a putative rare variant.* Using a BED file containing the loci of all putative rare variants, extract a subset of the original BAM file (using samtools), containing only reads overlapping these candidate regions. Convert these reads back to FASTQ format, and re-align using GSNAP [78]. GSNAP is much slower than STAR, but is considered one of the most accurate spliced-read aligners.

5. *Remove putative variants in genomic regions that cannot be uniquly mapped.* To minimize the chance of any possible misalignment, I also made sure that each candidate rare variant appears in a genomic region that can be uniquely mapped to the reference genome. To this end, I took a window of sequence upstream and downstream of the variant in question (for example, ± 50 bp), and aligned this sequence to the reference genome using BLAT. If there are many candidate hits elsewhere in the genome that are almost perfect matches to the queried sequence, I remove the variant from further analysis.

Misalignment is the dominant factor behind most false positives when calling RNA-seq rare variants. This is the reason why I decided to use three aligners as part of the new pipeline. I'm currently calling this pipeline the "three-pass" pipeline, but perhaps in the future I'll come up with a more clever name.

Most variant calling pipelines would not attempt to use STAR *and* GSNAP, and

BLAT is almost out of the question. On an average computer, STAR can process about 5,000 reads per second, GSNAP can process about 300 reads per second, and BLAT processes about 50 reads per second. Due to my large sample size, and limited computational resources, I decided to first align the reads using STAR (due to its exceptional speed), collect all putative rare variants, extract only the reads overlapping candidate variant regions (cutting down the size of the fastq file by at least a factor of 10, up to 100), re-align the reads using GSNAP, re-run the variant caller, find a new set of candidate variant regions, and determine the region's mappability using BLAT. Figure 2.4 outlines the software pipeline in more detail.

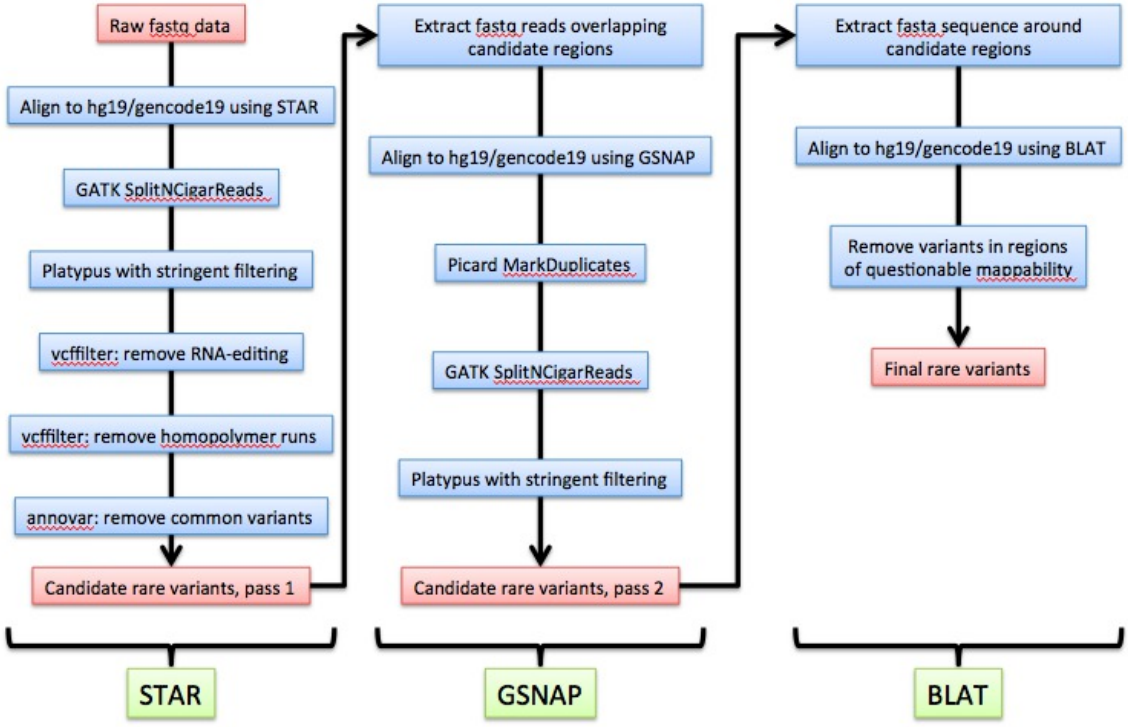

Figure 2.4: Software pipeline for confident calling of rare variants in RNA-seq data.

I first take the raw fastq files and align them to the hg19 human reference annotated with GENCODE version 19 gene annotations. I then sort and index the resulting BAM file using *samtools*. With the sorted file, I use GATK's *SplitNCigarReads* to split up junction-spanning reads. I re-sort and re-index with *samtools*, and then run *Platypus* to get an initial set of candidate variants. With the resulting VCF file, I remove RNA-editing instances using a BED file containing previously-discovered RNA-editing sites from RADAR and DARNED. I also remove variants in homopolymer regions with a BED file containing all known homopolymer run sites in hg19 (as discovered by repeatmasker). These removal steps are done with with the *vcffilter* tool in *VCFtools*. I then convert the final VCF file into an input format compatible with ANNOVAR (using one of annovar's scripts), and then run ANNOVAR, removing common variants with minor allele frequency $> 1\%$ in any observed subpopulation. The output is then converted into a BED file of all putative rare variants.

With this BED file, I use *samtools* to extract reads overlapping candidate rare variant regions. These reads are converted into FASTQ format, and then re-aligned to hg19 (with GENCODE V19 annotations) using GSNAP. Duplicates are then marked using Picard *MarkDuplicates*. The reads are again split using GATK's *SplitNCigarReads*, and *Platypus* is again used to call the variants.

For each candidate variant in the VCF file, the genomic sequence surrounding the variant of interest is extracted from the hg19 reference and converted into a

FASTA file. These sequences are then aligned by BLAT against hg19. Candidate regions that have alternative mapping sites with a quality score difference less than 2 are removed. In other words, if the region around a candidate rare variant can be mapped to another region of hg19 with almost equal mapping score, the variant is thrown out as a possible false positive. After this final step, the list of final candidates is outputted as a tab delimited file.

2.7 Evaluation of new RNA-seq rare variant calling pipeline

In order to evaluate the performance of this new "three-pass" pipeline, I ran the pipeline on the same data I used to evaluate the "regular" rare variant calling pipeline (the GIAB NA12878 WGS data and the ENCODE NA12878 RNA-seq data). The three-pass pipeline, compared to the standard pipeline, has significantly fewer false positives, cutting down the false positive rate from approximately 50% to about 5%. This reduction comes at a significant hit to sensitivity; about 50% fewer rare variants are called using the three-pass method.

Chapter 3

Genomic landscape of cancer

The previous chapter discussed some of the methods behind calling rare variants. In this chapter, I will go over some of the practical applications of calling variants, in relation to the field of cancer research. Before I do so, I will briefly go over some of the key concepts of cancer, discuss general categories of variants in cancer, and go over some of the simple heuristics that can evaluate a gene's role in cancer.

3.1 Tumor suppressors & oncogenes

Cancer is, at its core, a disease caused by the regulatory breakdown of cellular growth. There are many mechanisms that can be perturbed that can lead to this breakdown, such as amplification of growth signals, resistance to anti-growth reg-

ulators, resistance to programmed cell death (apoptosis), resistance to senescence, upregulated angiogenesis, development of metastatic behavior, evasion of the immune system, activation of chronic inflammation, and deregulation of metabolic pathways [79, 80].

Genes that contribute to cancer can be roughly placed into two categories: tumor suppressors and oncogenes. These two categories are often analogized as "brake pedals" and "gas pedals." Tumor suppressors restrict cellular growth, acting as checkpoints during cellular replication, detecting and repairing DNA damage, and promoting apoptosis [81]. Oncogenes typically involve genes that promote cellular growth, including transcription factors, growth factors, growth factor receptors, chromatin remodelers, signal transducers, and apoptotic regulators [82].

3.2 Drivers and passengers

Alterations in the DNA of a normal cell is what causes a normally-functioning gene to transform into a knocked-out tumor suppressor or a functional oncogene. Because a tumor suppressor gene acts to *prevent* cancer progression, mutations in these genes tend to "knock-out" gene function. Often times, even a single functioning copy of the gene is still functional enough to prevent cancer progression, hence the formation of the "two-hit hypothesis," that both copies of the gene need

to be knocked-out in order to be cancer-promoting [83].

Because oncogenes act to accelerate cancer progression, causative mutations in proto-oncogenes do not need to occur in both alleles, since a single affected copy will be sufficient in promoting cellular growth. Therefore, mutations in oncogenes tend to be "activating" or "gain-of-function" mutations, either by increasing protein expression, or by altering protein function (for example, by making a growth factor receptor constitutively active).

The car analogy can be extended to cover this phenomenon: Imagine a car with two brakes and two gas pedals (humans being a diploid organism). The car will still be stoppable with one failed brake pedal, as long as the other one still functions. However, the car only needs one gas pedal to be stuck in order to accelerate uncontrollably.

According to the Cancer Gene Census, over 1% of all genes are implicated in cancer [84]. Of the 572 genes currently implicated in cancer, about 90% have cataloged somatic mutations, 20% have known germline mutations, and 10% have both somatic and germline mutations [85].

There are a wide variety of cancer-promoting variants (driver mutations); they range from gross chromosomal changes to single point mutations. Approximately 95% of driver mutations are SNVs [86]. According to the Catalogue of Somatic Mutations in Cancer (COSMIC) version v70, there are approximately 2 million coding point mutations, 10,500 gene fusions, 60,000 genomic rearrangements, and

695,000 abnormal CNVs [87]. There may be a bias towards detection of coding missense mutations (due to an emphasis on exome sequencing), similar to the observational bias towards detecting massive exoplanets by using the transit method or the radial velocity method [88].

3.3 Development of a simple heuristic for classifying tumor suppressors & oncogenes

In order to gain a better understanding of which genes act as tumor suppressors, and which genes act as oncogenes, I worked with Dr. Bert Vogelstein to develop a simple metric for classifying genes in the Catalogue of Somatic Mutations in Cancer (COSMIC). Vogelstein noticed two key observations regarding COSMIC genes:

- *Mutation clustering in oncogenes.* For genes previously implicated as oncogenes, missense mutations tend to exhibit a large degree of positional clustering. For example, for KRAS, over 99% of cancer somatic mutations in KRAS are missense mutations, and almost exclusively affect residues 12 and 13 (Figure 3.1).

- *Enrichment for loss-of-function mutations in tumor suppressors.* For genes previously identified as tumor suppressors, there is a significantly higher pro-

portion of loss-of-function mutations. For example, over 90% of all cancer mutations observed in APC are either nonsense mutations or frameshift mutations (Figure 3.2).

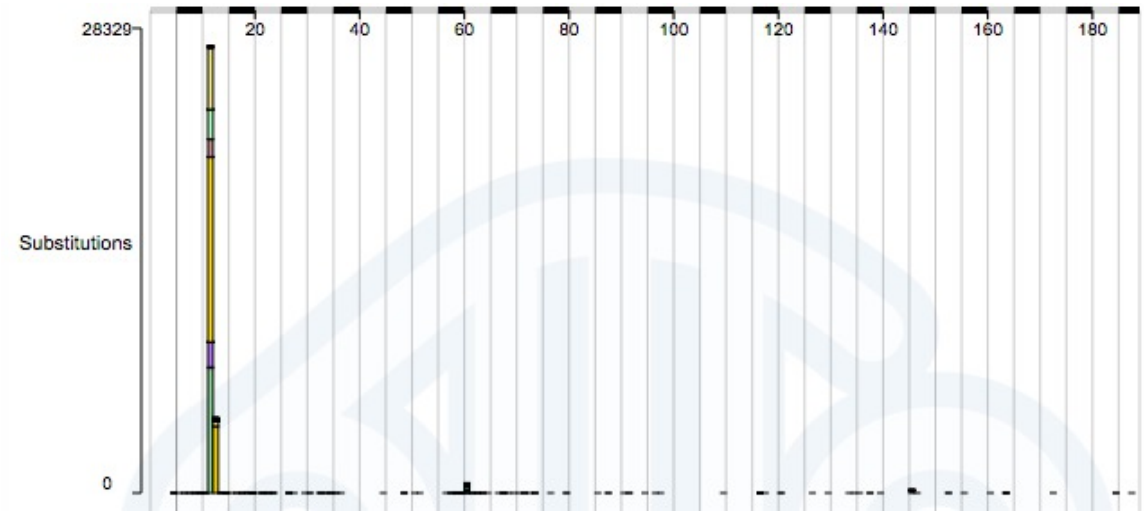

Figure 3.1: Histogram of KRAS missense mutations in COSMIC. Figure generated from the COSMIC website [89].

Color	Mutation Type	Mutant samples	Percentage
	Substitution nonsense	1880	42.12
	Substitution missense	756	16.94
	Substitution synonymous	261	5.85
	Insertion inframe	3	0.07
	Insertion frameshift	706	15.82
	Deletion inframe	2	0.04
	Deletion frameshift	1444	32.35
	Complex	21	0.47
	Other	93	2.08
	Total	4463	100

Figure 3.2: Distribution of mutation types of APC in COSMIC. Figure generated from the COSMIC website [89].

With these two key observations, Vogelstein and I created a simple decision-tree for classifying genes in COSMIC. For genes with more than 5 records in

COSMIC, if more than 15% of all records affect the same residue (and thus evidence of clustering), we classify that gene as an oncogene. If more than 15% of all records are loss-of-function mutations (frameshift or nonsense mutations), we classify the gene as a tumor suppressor (Figure 3.3). I named this heuristic "BUSDrivers," short for "*Bert's Updating Selection of Drivers*".

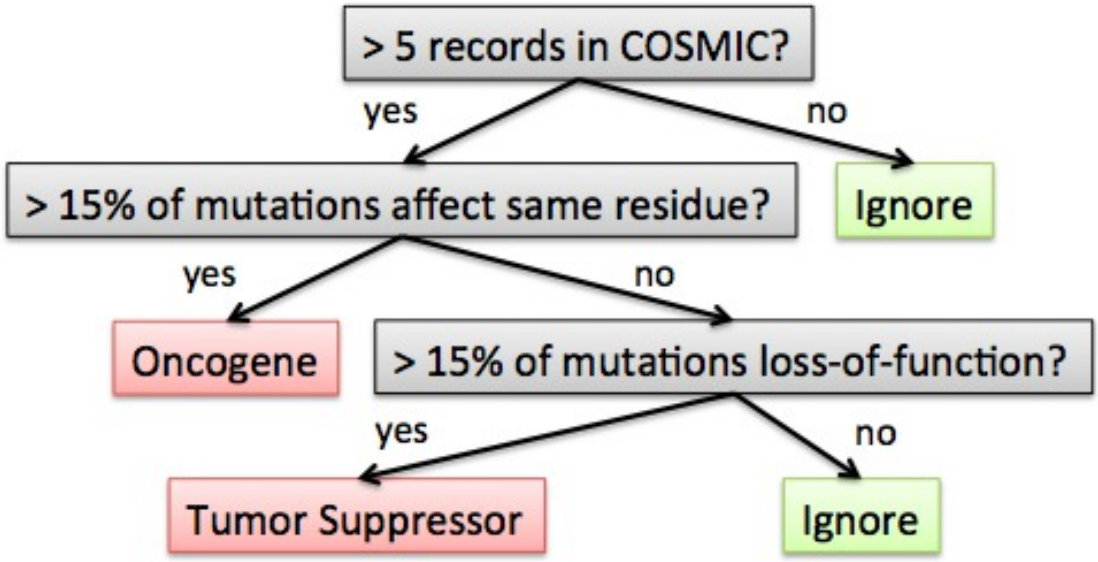

Figure 3.3: Decision tree for classifying genes in COSMIC.

Using *BUSdrivers*, we are able to correctly categorize many of the commonly-known tumor suppresors and oncogenes. BRAF, KRAS, EGRF, KIT, PIK3CA, and IDH1 are among the list of correctly-identified oncogenes. APC, NF2, WT1, and RB1 are among the list of correctly-identified tumor suppressors.

We also noticed a number of genes that fulfilled both conditions; more than 15% of mutations occur in the same amino acid, and greater than 15% of mu-

tations are inactivating. Among the list of genes that fulfill both these criteria are TP53, PTEN, VHL, NOTCH1, CDKN2A, FBXW7, RUNX1, and ATM. Most of these genes are typically considered tumor suppressors, but they also harbor gain-of-function mutations that act in a dominant negative fashion [90–95]. In other words, these point mutations are "so bad," they interfere with the normal functioning copy, subverting the two-hit hypothesis and making the tumor suppressor behave like an oncogene (in that it only required a single gain-of-function mutation in one allele).

3.4 Detection of different clustering signatures in genes implicated in autosomal dominant and autosomal recessive diseases

One of the main observations from the previous section is that in oncogenes, cancer-promoting mutations tend to be gain-of-function mutations, and they exhibit tight positional clustering. We wanted to extend this hypothesis, and see if genes implicated in autosomal dominant mendelian diseases might also present tight clustering of rare missense mutations. This section is based on material from

Turner *et al.* 2015 [96].

For autosomal dominant (AD) mendelian disease, only one copy of an affected allele needs to exist in order for disease to occur. This is in contrast to autosomal recessive (AR) mendelian disease, where a hit has to occur in both alleles. With this in mind, we wanted to explore whether AD and AR disease-related genes have missense mutational distributions similar to oncogenes and tumor suppressors. For an AD gene enriched with gain-of-function (GOF) missense mutations, we would expect these mutations to cluster around functional regions. For an (AR) gene, we may also expect loss-of-function (LOF) missense mutations to preferentially occur in functional regions.

We define clustering as missense mutation positions in a given gene occurring closer to each other (in the primary sequence) than would be observed by chance. In order to quantify clustering, I define a new, unbiased metric, CLUstering of Mutation Positions (CLUMP), which applies the Partitioning Around Medoids (PAM) clustering algorithm to a list of integer-indexed amino acid residue positions [97]. We use the *pamk* implementation of PAM *fpc* package in R. The number of clusters k is not specified in advance, but is estimated by varying k over multiple PAM runs and selecting k^* that yields the maximum average silhouette width. Thus, both the number of clusters and a "medoid," or representative member of each cluster are estimated by the algorithm. Next, for each cluster i, we compute the distance between each member of the cluster and its medoid and take a log

sum of these distances over all clusters. The final CLUMP score S_p for a given protein p is then defined as:

$$S_p = \sum_{i=1}^{k^*} \sum_{j=1}^{n_i} \frac{ln\left(\left|X_{ij} - m_j\right| + 1\right)}{n_i} \tag{3.1}$$

where X_{ij} is the position of mutation j in cluster i, m_i is the position of the medoid of cluster i, n_i is the number of mutations in cluster i, and k^* is the total number of clusters in the gene. The maximum clustering possible is when all observed mutations in all clusters occur in the same position as the cluster medoid, yielding a score of 0. In general, a protein with highly-localized mutations will have a low CLUMP score, while a protein with mutations spread across its protein sequence will have a high score.

To assess the statistical significance of S_p, we calculate a CLUMP score for all disease-associated missense mutations in a gene, and a separate CLUMP score for missense variants from 1000 genomes. Disease-associated missense mutations were taken from the Human Gene Mutation Database version 2014.2. To determine whether a gene is associated with AD or AR disease, we parsed abstracts in PubMed, looking for the keywords "autosomal dominant" or "autosomal recessive." We also filtered out any pubmed entries with an association to cancer, as we were only interested in mendelian disease. In order to get a high-confidence set of gene/disease associations, we disregarded genes with fewer than 12 publications, and required greater than 75% consensus towards AD/AR assignment. The

remaining 706 gene/disease associations were then manually checked for correct assignment using OMIM and primary literature search.

Using the Wilcoxon rank sum test, we find that proteins with AD mutations exhibited tighter clustering compared to 1000 genome variants ($P = 3.1 \times 10^{-12}$), and compared to AR proteins ($P = 8.4 \times 10^{-4}$). We also observe that AR proteins exhibit more localization than 1000 genomes as well ($P = 0.03$). A cumulative distribution plot of the CLUMP scores can be seen in Figure 3.4.

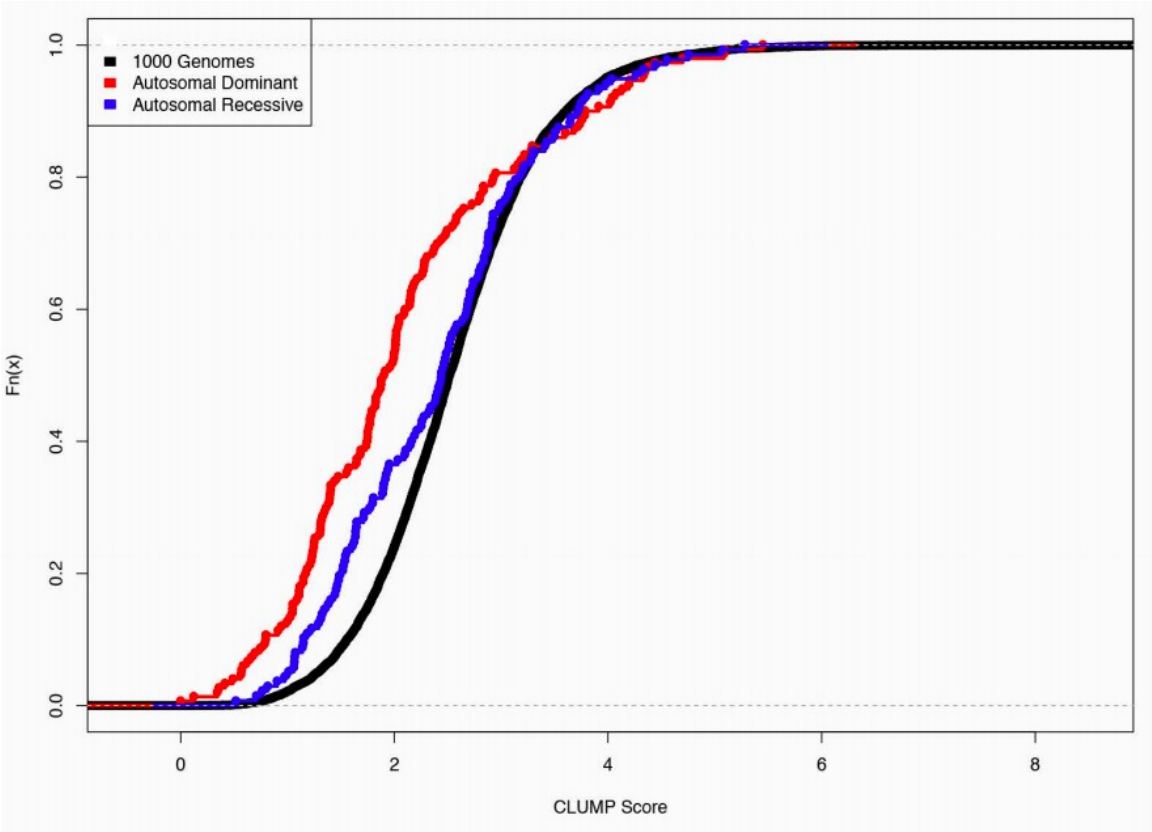

Figure 3.4: Cumulative distributions of CLUMP scores per gene

3.5 Classification of novel missense mutations in cancer

Cancer cells tend to accumulate many variants over time, but only a fraction of these variants will be cancer promoting. [98, 99]. Variants that promote tumor growth are classified as "drivers," while neutral variations that are just "along for the ride" are classified as "passengers." Determining whether a given variant is a driver or a passenger is an ongoing challenge.

Some types of variants are easier to classify than others. A stop mutation or a frameshift mutation has obvious impacts on the functionality of a gene. If the gene is previously known to be a tumor suppressor, we can safely assume an obvious loss-of-function mutation to be a driver.

Unfortunately, the majority of somatic mutations in cancer are single nucleotide variants, leading to missense variations in the amino acid sequence. Determining the functional impact of a single amino acid change is a much harder problem to solve.

With a good training set of known driver mutations and known passengers, supervised machine learning algorithms can be used to predict the functionality of a previously unencountered somatic missense variation. The most popular tool for classifying somatic missense mutations in cancer is CHASM, short for Cancer-specific High-Throughput Annotation of Somatic Mutations [100]. CHASM uses a

random forest trained on 3,299 driver missense mutations from COSMIC, curated by Bert Vogelstein, and a set of 4,500 synthetically-generated passenger mutations. The passenger mutations are generated by modeling the tumor mutation background using the dinucleotide contexts. For each mutation, a set of 70+ features are annotated, ranging from features such as multiple sequence alignment (MSA) entropy, to amino acid changes (e.g., change in hydrophobicity), to exonic SNP density. The random forest uses an ensemble of decision trees, each trained on a random sampling of the data and the features. To evaluate a novel missense variant, each tree in the forest casts a "vote," as to whether the features in the novel variant are simliar to that of a driver or passenger mutation.

Chapter 4

Database of precomputed cancer-related features in the human exome

For the purposes of illustration, let us assume we wish to create a machine learning algorithm that can, whenever we take a photo, recognize whether the image contains a cat. The performance of such an algorithm (and all supervised machine learning algorithms), depends on the following factors:

- *Clean training data.* Care must be taken to ensure that the labels for each training datapoint are correctly assigned. If 10% of our training data cat pictures are actually pictures of dogs, then our classifier is no longer a "cat pic" classifier, it becomes a "fairly good cat, but occasionally a dog" pic

classifier.

- *Training data is a representative sampling of data to be evaluated.* It is critical to have a clear understanding of which data are in the "positive" class, and which kind of data are in the "negative" class. If our positive training data are pictures of cats, and our negative training data pictures of dogs, then our classifier is meaningless if we present it with ostrich pictures for classification.

- *Informative features.* Feature engineering is crucial for good classifier performance. Although most classifiers have some degree of robustness when it comes to ignoring useless features, there must at least be a handful of truly informative features that can discriminate between the classes. Features such as the timestamp (when the photo was taken), intensity of the 17th pixel, number of legs, would probably be useless when it comes to evaluating cat and dog pictures.

In the last section of the previous chapter, I went over some of the basic workings of CHASM, a random forest classifier trained to discriminate driver mutations from passengers. Some of the key aspects behind CHASM are its 70+ features used to annotate a candidate missense variant.

Feature annotation tends to be the most time- and data-intensive steps when using machine learning. In this chapter I will go over some of our work involving

a database of pre-computed features covering all possible nucleotide changes in the annotated human exome. This database can either be used in a stand-alone fashion, or as part of a larger variant discovery / variant annotation pipeline. This chapter is based on material from Wong *et al.* 2011 [101].

SNVBox, the name of our precomputed feature database, is a MySQL database of 86 predictive features relevant to the biological impact of a SNV. The features have been pre-computed for each codon in all protein-coding exons of annotated human mRNA transcripts in the NCBI RefSeq, CCDS, and EBI Ensembl databases [102–104]. A visualization of the database schema can be seen in Figure 4.1, and a description of available tables in the SNVBox database is shown in Table 4.1. The `SnvGet` program, included as part of the SNVBox package, allows for fast retrieval of selected features from the SNVBox database for classifier training and for scoring of mutations inputted by the user.

SNVBox comes packaged with CHASM, which is an open-sourced collection of Python and C++ programs that can take a list of somatic missense mutations as input and rank them according to their likely tumorigenic impact. Users have the option to use their own estimates of passenger variant frequencies to generate synthetic passenger mutations, or select from a library of pre-generated passenger frequency tables derived from several common cancers.

PyInstaller version 1.4 was used to package the Python source code into dynamically linked, executable binaries. The `SnvGet`, `BuildClassifier`, and `RunChasm`

executables are run by the user on the command line. The statically-compiled C++ executable `waffles_learn` from the WAFFLES machine learning library is also used internally.

4.1 Program workflow

1. Prepare an input file of estimated passenger mutation rates in the cancer of interest. Optionally, select from one of several pre-computed passenger rate tables.

2. Prepare an input file of missense SNVs to be classified. Each row contains a protein accession identifier, codon number, and reference and variant amino acid residues.

3. Run the `BuildClassifier` program.

    - `BuildClassifier` generates *in silico* passenger mutations by introducing random nucleotide substitutions in a sequence library of expressed human mRNA transcripts from NCBI RefSeq, according to the distributions specified by the passenger mutation rate table.

    - `BuildClassifier` fetches a list of predictive features for each variant in the training set from SNVBox.

    - `BuildClassifier` then trains the random forest classifier using `waffles_learn`.

4. Execute the `RunChasm` program.

    - `RunChasm` retrieves features from SNVBox, for all inputted variants.
    - CHASM then scores each candidate variant supplied by the user.
    - A second set of passenger variants are generated, filtering out variants in genes that have been previously implicated in cancer. These genes are taken from COSMIC, the Cancer Gene Census, and all cancer gene-sets in MSigDB [84, 105].
    - These filtered passengers are then scored by CHASM, generating an empirical null of variant scores.
    - This null distribution is then used to calculate the P-value for each user-submitted variant (by calculating the fraction of the null distribution having CHASM scores equal to or more extreme than the variant CHASM score).
    - The P-values are then adjusted using the Benjamini-Hochberg multiple testing correction [106].
    - A tab-separated list of user-supplied variants, CHASM scores, P-values, and Benjamini-Hochberg estimated false discovery rate (FDR) is finally outputted.
    - An attribute-relation file format (ARFF) file is also outputted, containing model data in a format optimized for some machine learning pack-

ages.

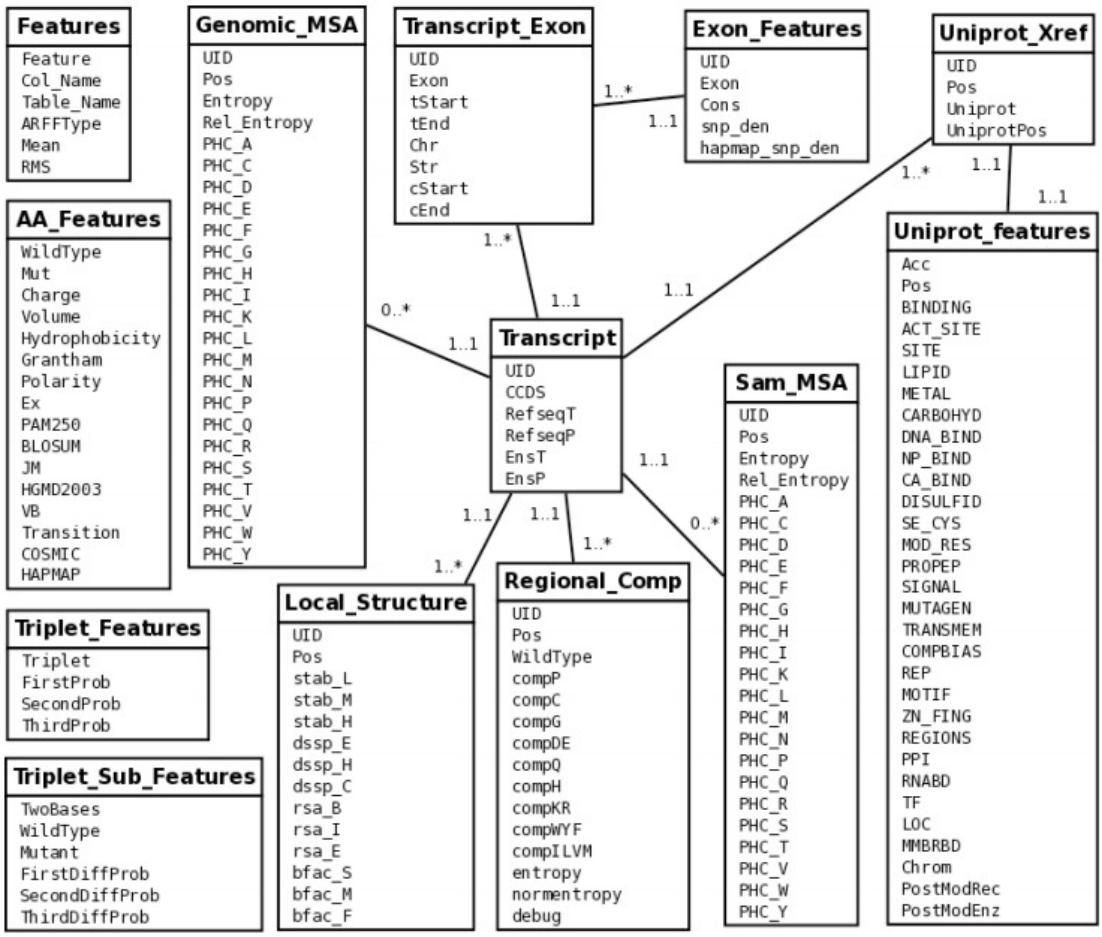

Figure 4.1: SNVBox database schema.

4.2 SNVBox discussion

The CHASM/SNVBox toolkit is the first distributable software package that specifically targets somatic missense mutations in cancer. The random forest classifier in CHASM is tuned to discriminate between known drivers and a set of

Table 4.1: Description of tables in SNVBox.

SNVBox Table	Description	Example features
AA_features	Amino acid substitution features.	Volume change, hydrophobicity change, frequency of change in COSMIC
Exon_features	Exon conservation and variability	46-way exon conservation, SNP density
Features	Metadata table of features	SNVBox feature name, table name, column name
Genomic_MSA	46-way vertebrate conservation	Relative entropy, absolute entropy, PHC scores
Local_Structure	Predicted protein structure	Stability, predicted helix, predicted coil, solvent accessibility
Regional_Comp	Regional amino acid composition	Proportion of glycines, prolines, AA entropy
Sam_MSA	Conservation calculated with SAM-t2k HMM MSA	Entropy, relative entropy, PHC scores
Transcript	mRNA accession IDs in CCDS, RefSeq, and Ensembl	RefSeq transcript ID, RefSeq protein ID, CCDS accession
Transcript_Exon	Exon boundaries in each mRNA transcript	Exon ID, chromosome, start, stop, direction of transcription
Triplet_Features	Triplet AA frequencies calculated by SwissProt	First position probability, second position probability
Uniprot_Xref	Cross-reference mapping protein position to UniProt position	Protein ID, protein position, UniProt ID, UniProt position
Uniprot_Features	Protein features derived from Uniprot	Binding site, active site, metal binding site, disulfide bridge, zinc finger region

synthetically generated passenger mutations that match the mutational spectrum of the cancer of interest. CHASM results are sensitive to this mutational spectrum, and users are encouraged to use the somatic variant calls from their sequencing data to create the best possible estimate of the mutational spectrum.

While many SNV classifiers are available via web interfaces, these tools are not capable of handling large input datasets (e.g., thousands to millions of SNVs discovered through high-throughput sequencing projects). Researchers have developed distributable packages that users can run on their local system to enable high-throughput SNV processing. These packages depend on third-party databases (sequences, alignments, protein structures, specialized protein annotations) and third-party software packages. The popular PolyPhen package, for example, requires installation of 10 third-party software packages, in addition to three Perl modules. To our knowledge, all available SNV classification tools must calculate or retrieve features "on-the-fly," almost always using third-party databases and software in the process. In contrast, the predictive features available in SNVBox (also calculated by third-party tools) have been exhaustively pre-computed, allowing rapid retrieval for any given custom dataset. In benchmark testing, retrieval of 86 features for one million SNVs took 11.39h on a Dell R900 server with two AMD Opteron dual-core 64-bit CPUs with 16GB of RAM. CHASM score computation for these one million SNVs took an additional 10m and 33s.

Finally, the predictive features available in SNVBox have been designed to be useful for classification of both germline and somatic SNVs. We hope that SNV will enable design of new, improved machine learning algorithms to predict the impact of SNVs.

Chapter 5

Database mapping positions & features from 1D sequence position to 3D protein structure

With the advent of cheap next-gen sequencing, it is now possible to execute large-scale studies generating millions of SNVs of interest. Bioinformatic methods have been developed to predict a given SNV's impact on protein function, allowing for downstream prioritization of high-scoring SNVs. Interpreting a high-scoring SNV's impact on actual protein structure or function remains a challenge. One approach for interpreting SNV function is to map the SNV onto the three-dimensional protein structure in which it occurs.

Working with SNVs in the three-dimensional context of the protein confers

two advantages:

- *Variant interpretability.* By being able to visualize the variant in the context of the 3D protein environment, we can more easily create hypotheses on how the SNV affects protein function or structure.

- *More informative distance metric.* Proteins are three-dimensional creatures. What may be distant in one-dimensional sequence space may be extremely close in three-dimensional space due to protein folding. A more meaningful distance metric will make downstream analysis more accurate, especially when trying to quantify clustering.

In this chapter, I will go over the details of a database I created that maps 1D sequence position (and relevant features) to available 3D protein structures. We named this tool MuPIT-DB (Mutation Position Imaging Tool Database). This chapter is based on material from Niknafs *et al.* 2013 [107].

The MuPIT database is the back-end to the MuPIT visualization tool, a web-server written in Java. The Java servlet encapsulates a series of Django applications: A Jmol applet for 3D protein visualization, a database (DB) interfacer accessing the MuPIT DB, and an overview page constructor [108]. Javascript functions are used to allow for user interactions between the overview tables and the 3D visualization.

The MuPIT database contains mappings from genomic coordinates to PDB

structure coordinates, using methods developed in the LS-SNP/PDB application [109]. In brief, a software pipeline uses BLAT to align protein sequences from Protein Data Bank (PDB) structures to human protein sequences in the UniProt knowledgebase (UniProtKDB) [77, 110, 111]. UniProtKB protein sequences are aligned to RefSeq mRNA transcript sequences with tBLASTn, and mRNA transcript sequences are aligned to human genomic DNA with BLAT [112]. Currently, only the canonical isoform of each UniProtKB protein sequence is supported. PDB sequence is taken from a local mirror of the PDB database [110]. The UniProt sequences and feature annotations are taken from a local mirror of the UniProtKB [111]. Alignments are calculated on a high-performance computing cluster, managed by the Son of Grid Engine (SGE) batch-queuing system. The resulting annotations and alignments are stored in a MySQL database, which is updated on a monthly basis to ensure synchrony with its external sources.

5.1 Coverage

MuPIT currently provides at least one viewable PDB structure for 3,641 human proteins (which is 18% of the 20,226 SwissProt-curated proteins in UniProtKB), and covers 924,857 codons in protein-coding exons (which is 8.2% of the 11,291,814 total codons in SwissProt). Coverage of cancer-related genes is somewhat higher: 45% of 487 genes in the Cancer Gene Census (March 2012 ver-

sion) and 16% of the codons in these genes are covered in the current version of MuPIT [84].

5.2 Database pipeline

The Mupit DB can be divided into three main parts:

1. Database of aligned SwissProt position to PDB position,
2. Database of aligned PDB position to genomic coordinate,
3. Database of features mapped to PDB position.

5.2.1 SwissProt to PDB alignment

The SwissProt-to-PDB alignment begins by calling helper functions to fetch the relevant SwissProt and PDB sequences. Pairwise alignment is done using BLAT, and the resulting PSL file is then evaluated using a Python script (`filterPSL`). If the alignment is not redundant and of sufficient quality, the PSL is stored into the MySQL database (into the table `Sprot2PDBAlign`), and positional mappings are calculated using another helper function (`pdb2sprot`). These mappings are then placed into the `Sprot2PDB` MySQL table (Figure 5.1).

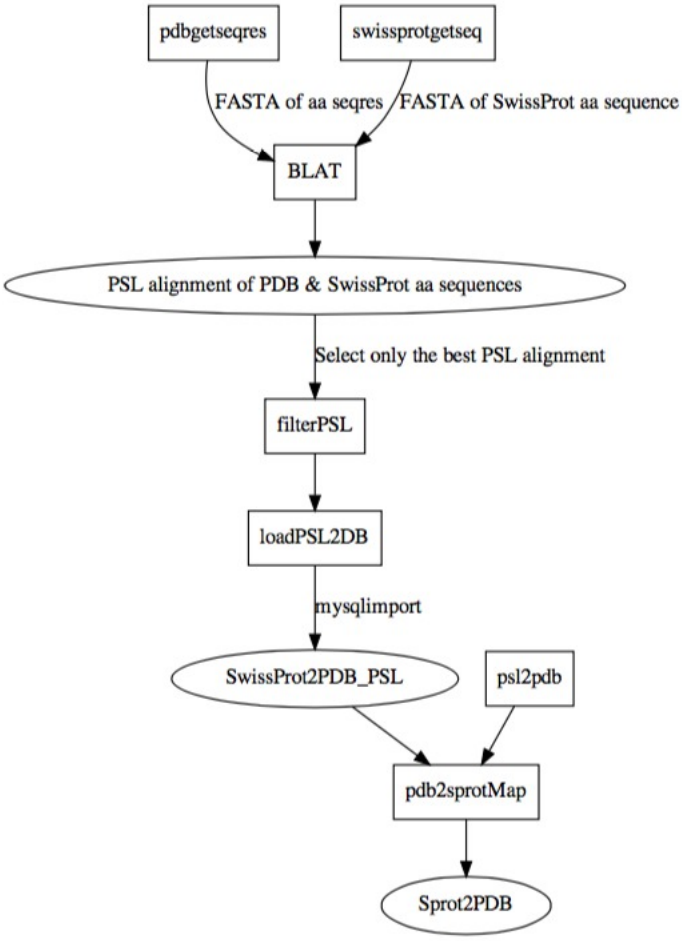

Figure 5.1: Detailed flowchart for mapping SwissProt sequences to PDB sequence.

5.2.2 PDB to genomic coordinates

In order to convert PDB positions to chromosomal coordinates, we first map the SwissProt sequences onto RefSeq, and then fetch the RefSeq-to-genome alignments. The SwissProt/RefSeq alignments are done using the script `sprot2Refseq`, which is a wrapper around tBLASTn, taking in a protein sequence, and mapping them onto matching nucleotide sequences. The resulting blastXml file is then con-

verted into a PSL file, and filtered using `filterPSL`. Any alignments that pass the filter are then stored into the `Sprot2RefseqAlign` table. RefSeq-to-genome alignments are fetched using the `refSeqGet` tool from the Kent source utilities, and then placed into the `Refseq2GenomeAlign` table [113]. Once the alignment tables are in place, another python helper function `genome2pdbMap` is used to get the final PDB to genome mappings, which are then placed into the `Genome2PDB` MySQL table.

5.3 Mapping SwissProt features onto PDB structures

Once the `Sprot2PDB` table has finished populating, we can use this table to map the SwissProt features (which populate the `SwissProt_Features` table) onto matching PDBs. The helper function `sprot_features_full` will execute the mappings, and place the results into `PDB_Features`.

Chapter 6

Analysis of rare variants in schizophrenia

6.1 Genetics of schizophrenia

Schizophrenia is one of the most prevalent psychiatric disorders, presenting a lifetime risk of 0.7% worldwide [114]. Genetics have been shown to be the most dominant risk factor behind schizophrenia; twin and family studies have shown a heritability of $\sim$ 70%, one of the highest among psychiatric diseases [115–117]. In spite of the high heritability, the genetic mechanisms driving schizophrenia are still unclear.

Several large-scale studies have been developed to determine whether schizophrenia falls under a set of commonly-encountered genetic paradigms. Because schizophre-

nia is a relatively common disease, genome-wide association studies (GWAS) have been carried out to determine whether the common disease-common variant hypothesis explains schizophrenia [118–121]. GWAS studies have discovered many significant SNPs and implicated several genes, but they have not discovered any "smoking guns." There doesn't seem to be any single SNP, or group of SNPs, that can fully account for schizophrenia. S. Hong Lee *et al.* have estimated that SNPs account for 23% of variation in liability to schizophrenia [122].

Another popular paradigm is the common disease-rare variant hypothesis (CDRV). A small set of highly penetrant rare variants may affect one or more genes, leading to disease. Rare variants are typically defined as a variant with a minor allele frequency less than 0.5% or 1%. Several whole exome sequencing studies on schizophrenia have suggested a polygenic burden of rare variants implicated in schizophrenia, but with the possible exception of TAF13, no single gene is shown to be significantly enriched in rare variants in schizophrenics [123–126].

In light of these promising results from GWAS and exome sequencing studies, a common reaction is to increase the sample size in the hopes of uncovering more causative genes/variants leading to schizophrenia. This is an effort to increase *statistical power*. Statistical power is defined to be the probability of rejecting the null hypothesis H_0 provided that the alternative hypothesis H_1 is actually true. There are five main factors that influence statistical power:

- *Measurement error.* Increasing precision (reducing the measurement error) for each sample reduces the variance, making it easier to discriminate a bimodal distribution.

- *Effect size.* A large effect size (e.g., standardized mean difference between two populations) usually leads to an increased distance between peaks of a bimodal distribution, also facilitating discrimination.

- *P-value cutoff.* The significance cutoff controls the α, the type I error, and changing this cutoff also indirectly influences the type II error (β), which is the complement of power ($1 - \beta$).

- *Study design.* Focusing on samples lying on the extreme ends of a distribution of interest, or relying on a family with multiple affected phenotypes can increase the likelihood of discovering significance.

- *Sample size.* Increasing the sample size creates a better estimate of the population distribution, increasing the likelihood of discovering small differences between distributions.

Conducting future studies of identical design with increased sample size is a double-edged sword. Statistical power will be increased, but often with a hit to the effect size of observed significant results. If there is a gene/variant with a huge effect size, there would be no need to conduct a larger study, since pre-

vious studies with smaller sample sizes should have been able to detect these results [127,128].

6.2 Non-coding rare variants

In light of the observations made from previous GWAS and whole exome sequencing studies, I wanted to see if it was possible to examine areas of the genome that were previously overlooked. GWAS can pick up common variants in coding, noncoding, and intergenic regions. Exome sequencing studies are primarily focused on coding variants. In this section, I will go over some of the reasoning behind looking for rare variants in noncoding regions.

Once overlooked in favor of protein-coding genes, non-coding RNA (ncRNA) have started to enter the research limelight, thanks to recent whole transcriptome sequencing efforts that highlight the ubiquity of noncoding RNA in the genome [129]. Non-coding RNAs have also been implicated in brain development, synaptic plasticity, and psychiatric disease [130,131]. The long non-coding RNA (lncRNA) Gomafu has been shown to be dysregulated in schizophrenia [132].

Introns have also been overlooked in the past, thought to be little more than a transcriptional speedbump on the path to a translated protein. Due to the fact that most introns are spliced out of the final transcript, it can be easy to assume that most introns would have little to no effect on gene function. The following is

a list of the ways an intron can affect gene function; this list is taken from Chorev *et al.* 2012 [133]:

- *Transcription initiation.* Enhancers and silencers in the 5′-most intron can regulate transcription initiation. Alternative promoters can also initiate transcription in a separate downstream location from the canonical transcription start site.

- *Genes inside introns.* Also called nested genes, these are transcripts that reside completely within the intron of a larger, encapsulating gene.

- *Delay of transcription.* Extremely long introns increase the overall length of transcription, causing the polymerase to take up to several extra hours to fully transcribe the gene.

- *Alternative splicing.* Splicing regulatory elements (SREs) are motifs inside the pre-mRNA of interest that recruit splicing factors that either enhance or silence alternative splicing. Such elements inside an exon are called exonic splicing enhancers and exonic splicing silencers (ESEs and ESSes), and such elements inside introns are called intronic splicing enhancers and intronic splicing silencers (ISEs and ISSes).

- *Nonsense mediated decay (NMD).* After splicing, an exon-junction complex (EJC) persists in the pre-mRNA, which acts as a signpost of past splicing.

Existence of EJCs significantly downstream from the stop codon triggers NMD.

- *Cytoplasmic localization.* EJCs also seem to play a role in cytoplasmic localization. For example, alternative isoforms of *BDNF* determine whether *BDNF* localizes in the dendrites or the soma of a neuron [134].

- *Intron retention.* Considered the least prevalent form of alternative splicing in mammals, a recent paper has shown possible evidence of widespread intron retention in human and mouse poly-A RNA-seq data [135].

- *Recursive splicing.* Instead of being removed as a single unit, long introns that contain recursive splice sites (RS-sites) allow for multiple splicing events to remove the long intron iteratively. In vertebrates, long introns contain a small recursive splicing exon (RS-exon) that allows for recursive splicing to occur. These RS-exons may remain in the final spliced product, affecting mRNA function [136].

Introns may be of particular interest in psychiatric disease. Brain-specific genes are longer in length, and tend to contain extremely long introns [137]. Brain-specific genes also seem to have developed mechanisms for regulating long genes; the gene MECP2 represses long genes by binding to methylated `CA` regions; disruptions in the MECP2 gene lead to Rett syndrome [138]. The gene TDP-43 is involved in alternative splicing regulation, and has been shown to sustain and

upregulate long pre-mRNAs by binding to GU-rich intronic sites; disruptions in this gene have been implicated in amyotrophic lateral sclerosis [139].

The different intronic effects on gene function also rely on different aspects of the intron. Some effects, such as those relying on the EJC, such as NMD and localization, depend merely on the existence of an intron, regardless of its length or sequence. Transcriptional delay mechanisms depend on the length of the intron — the longer the intron, the longer it takes for polymerase to transcribe [140].

Of special interest are the intronic regulators that depend on the actual intronic sequence. Intronic enhancers & silencers, intron retention, recursive splicing, and alternative splicing all can be pertubed by intronic variants. By using a deep neural network, Xiong *et al.* were able to establish that intronic variants can impact splicing performance; by applying their method to a whole-genome sequencing case/control dataset of autism, they demonstrated that most of the high-scoring variants are intronic [141].

In light of these observations of non-coding RNA and introns affecting neuronal regulation and neurological disorders, I thought it would be worthwhile to examine these regions in a case/control study of schizophrenia, since these regions have been overlooked in previous studies.

6.3 RNA-seq dataset

In order to poll the non-coding regions of the genome, I relied on Ribo-depleted (RiboZero Gold) RNA-seq data, which we have shown in a previous chapter to contain many reads mapping to intronic regions, and are capable of detecting noncoding RNA with higher sensitivity than poly-A RNA-seq.

Table 6.1 contains the number of post-mortem brain RNA-seq samples used in this chapter. Due to the low number of asian and hispanic samples, these samples were removed from further analysis.

Table 6.1: Number of control and schizophrenic RNA-seq postmortem brain samples.

Race	Condition	Number of samples
African american	Control	180
	Schizophrenic	85
Asian american	Control	6
	Schizophrenic	3
Caucasian	Control	133
	Schizophrenic	81
Hispanic	Control	9
	Schizophrenic	3

6.4 Quality-control analysis of RNA-seq samples

Before we do any serious analysis of the RNA-seq samples, it is important to determine if there are any samples that have quality control metric outliers — including these samples in downstream analysis might introduce unnecessary false positives, or decrease statistical power (due to the samples being too lousy for any sort of genomic variation detection). The dataset in Table 6.1 was chosen after manual curation of fastQC results. FastQC is not the be-all, end-all solution for QC, unfortunately. Read aligners also generate QC reports, with metrics such as *proportion of successfully mapped reads*, *proportion of reads that are spliced*, etc. *RSeQC* is also used to generate QC metrics of generated BAM files. In this section, I use both of these QC reports to look for any possible outliers or systematic biases in the RNA-seq data.

The STAR aligner was used to align reads to the GENCODE V19 annotated hg19 human reference [6, 35]. The proportion of reads successfully mapping to the GENCODE V19 hg19 was taken from the STAR QC logs. RSeQC was used to determine the proportion of reads mapping to the mitochondrial genome (chrM) and the proportion of ribosomal RNA (rRNA) reads [43].

The RNA integrity number (RIN) is a commonly-used metric for RNA quality. Before the introduction of the RIN value, RNA quality was often estimated

by considering the ratio of 28S and 18S ribosomal RNA, since 28S rRNA tend to degrade more quickly than 18S rRNA. This method has been shown to be inconsistent [142]. The RIN value incorporates several other features, including the ratio of 18S/28S reads vs all reads and the total amount of 28S reads. For this chapter, the RIN values were generated by an Agilent Bioanalyzer.

Figure 6.1 shows a scatterplot of the pair-wise relationships between the percentage of mapped reads (pmappedreads_max), the proportion of reads mapping to the mitochrondrial genome (p_chrm), the proportion of ribosomal RNA reads (p_rrna), and the RIN. From this diagram, it is clear that there are a significant number of outliers. Luckily, there does not seem to be a systematic bias between cases and controls in any of the metrics. There is a clear relationship between the total number of mapped reads and the proportion of reads mapping to chrM and rRNA.

I hope I'm not spoiling the ending here, or ruining the narrative, but in truth, I actually ran the variant calling pipeline first, without doing this QC first. In retrospect, I should have filtered out the samples first. In any case, using the variant calling pipeline, I was able to determine the number of high-confidence variants called in the data. What made me go back and check the QC results was the discovery that there were a handful of samples with an alarmingly low number of detected variants (Figure 6.2).

A usual QC cut-off is to set a threshold on the percentage of mapped reads. For

this study, I decided on an arbitrary cutoff of 70%, this seems to do a decent job getting rid of samples that have a pathologically low number of variants called.

6.4.1 Quality control of polyA RNA-seq samples

The previous section covered the QC I did on the RiboZero Gold RNA-seq samples. We also have a set of polyA RNA-seq samples from the same individuals, and I figured I might as well run a similar QC check on these samples, in case we wished to use this dataset for downstream analysis.

After selecting samples that passed fastQC checking, I first aligned the reads using STAR, and looked at the STAR QC metrics. Looking at the histogram of the percentage of mapped reads, it appears that there may be a possible bimodal distribution — a tight distribution of "high-quality" samples with percentage $>$ 90%, and a "flatter" distribution of samples with a large range of not-so-great mapped read percentages. This bimodal distribution is troubling, especially if the two distributions correlate with disease condition (Figure 6.3).

A strip-plot is a great way to get a quick eye-balling of outliers. In Figure 6.4, we notice that there are outliers for each disease class.

In order to highlight the importance of not relying on just the RIN, I created a joint plot of the percentage of mapped reads against the RIN (Figure 6.5). There does not seem to be a clear relationship between the two variables. Therefore, it is possible that a sample can have a great RIN, but awful mapped read percentage,

rendering the sample useless for downstream analysis. Running a similar analysis comparing RIN to the proportion of reads mapping to chrM (the mitochondrial genome) shows a strong relationship between the two metrics (Figure 6.6).

The aim of this subsection is to determine a list of samples that have outlier QC values. To get an idea of what the outliers look like, I created a grid of pairwise scatterplots of five QC metrics (Figure 6.7):

1. Percentage of reads successfully mapped to hg19
2. Average length of mapped read
3. Read mismatch rate
4. Proportion of reads mapping to multiple loci
5. RIN
6. Proportion of reads mapping to chrM

A 5x5 grid of scatterplots is an overwhelming visualization. In an effort to make the search for outliers easier, I decided to plot the first two components of the principle component analysis (PCA) of this data (Figure 6.8). Not only do I get a cleaner visualization of the QC data, this also helps when I run an algorithm to identify outliers, since I get to avoid the Curse of Dimensionality.

To algorithmically detect outliers, I use a one-class support vector machine (SVM), trained on the PCA-transformed QC data. The one-class SVM is an unsu-

pervised machine learning algorithm that attempts to create a higher-dimensional representation of the data, using a radial basis function. In other words, the algorithm tries to find the "center" of the datapoints in feature space, then draws a hypersphere around the datapoints, attempting to create the smallest possible hypersphere able to enclose as many datapoints as possible. Figure 6.9 shows the heatmap of the one-class SVM outlier predictions. In this way, we are able to identify 7 outliers that seem to significantly deviate from the rest of the samples. Plotting these seven outliers on the pairwise scatterplots confirm that these outliers are also outliers when considering each QC metric separately (Figure 6.10).

By running the one-class SVM outlier detection algorithm on the PCA-transformed QC metrics, we were able to identify a set of strange-behaving samples that we would not have isolated had we simply used a RIN cutoff, or a mapped read percentage cutoff. Of course, we will still use a RIN cutoff for downstream analysis, and if we're concerned about gene expression, a chrM proportion cutoff would also make sense (or at least some way of correcting the large number of chrM reads).

6.5 Calling and analyzing rare variants in the Ribo-depleted RNA-seq data

To reiterate — this chapter chronicles my attempt to analyze rare non-coding variants in a case/control study of schizophrenia. Now that we examined the QC results of our data, in this section, I will go over the computational pipeline I used to generate the list of rare variant candidates to consider.

After BAM files are generated by aligning fastq reads using STAR (on GENCODE V19 hg19), the resulting BAM files are sorted and indexed using SAMtools. These BAM files are then fed into the three-pass pipeline as described in section 2.6. The resulting candidate rare variants are prioritized using FATHMM-MKL [70]. Unlike most variant interpreters, FATHMM-MKL is capable of predicting the functional impact of noncoding variants. This functionality is achieved by using a training set of noncoding variants, and including features that are available in noncoding regions, such as the 46-way vertebrate multiple sequence alignment, GC content, and features in ENCODE, such as histone modification sites, open chromatin regions, and transcription factor (TF) binding sites.

Once the rare variants predicted by FATHMM-MKL to be functional are selected, I run a gene-burden test on each gene to determine whether the tested gene is enriched in rare variants. The resulting set of p-values are then corrected for multiple tests.

The subsequent subsections will go into more detail on the rare variant analysis pipeline.

6.5.1 Allele frequency cutoff

Although the three-pass pipeline uses ANNOVAR's "popfreq_max" table to determine common variants to filter-out, there may be some yet-to-be annotated common variants, or other common variants that have not been properly annotated by ANNOVAR. Also, systematic false positives called by the three-pass pipeline might also show up as a common variant. Figure 6.11 is an allele frequency histogram of the putative rare variants. Here, the number of putative rare variants are plotted as a function of the number of samples containing the variant. For example, there are approximately 10 putative rare variants that are detected in 300 samples. If ANNOVAR truly removed all known common variants, these 10 putative rare variants must be false positives, and should be removed from further analysis. To minimize the risk of introducing false positives (or possible common variants), I removed all variants that have a sample-wide minor allele frequency greater than 1%.

Table 6.2: Biotype distribution of putative RiboZero RNA-seq rare variants.

Variant biotype	Number of variants
downstream	13201
exonic	73921
exonic;splicing	44
intergenic	138349
intronic	1157463
ncRNA_exonic	19779
ncRNA_exonic;splicing	24
ncRNA_intronic	131742
ncRNA_splicing	61
splicing	294
upstream	4824
upstream;downstream	858
UTR3	90685
UTR5	11264
UTR5;UTR3	286

6.5.2 Biotype distribution of putative rare variants

After removing commonly-occurring variants from the list of putative rare variants, I used ANNOVAR to annotate the biotype of each variant. Table 6.2 lists the number of variants that fall into each biotype category. As expected, the number of intronic rare variants dominate all other categories. This may be an issue when it comes to interpreting the rare variants, as we can expect that most of these intronic rare variants may be neutral.

Table 6.3: Number of different types of rare variants detected in the RiboZero RNA-seq data.

Variant type	Number of variants
SNV	1475530
Insertion	48687
Deletion	118578

6.5.3 Prioritization of putative rare variants

Since most detected variants are intronic, and there is no expectation for most of these variants to be functional, it is essential to try and prioritize these variants. Otherwise, downstream statistical analyses will be polluted with meaningless variants, destroying statistical power.

I used FATHMM-MKL to prioritize my variants. FATHMM-MKL uses a support vector machine (SVM) with features transformed using multiple kernel learning (MKL). A support vector machine is a supervised machine learning algorithm that attempts to find a hyperplane that best separates the two classes in feature space. Classically, this classifier is linear (linear-SVM), but by transforming the feature space (the "kernel trick"), SVM is capable of non-linear classification [143]. Instead of manually choosing a kernel, FATHMM-MKL uses multiple kernel learning, which uses a combination of differently-weighted kernels to transform the data [144]. MKL takes the guesswork out of manually choosing a static kernel function, and can be advantageous when incorporating heterogeneous features, since the optimal kernel can be chosen for each feature type.

FATHMM-MKL can only handle single nucleotide variants, so I removed 167,265 insertions and deletions from the list of candidate rare variants, and allowed FATHMM-MKL to prioritize 1,475,530 SNVs. FATHMM-MKL classifies about 56% of exonic rare variants and 5.4% of intronic variants to be deleterious (Figures 6.12 and 6.13).

6.6 Case/control analysis of rare variants in schizophrenia

Single variant tests are feasible with common variants, but for rare variants, it is extremely difficult, if not impossible, to see if a singular rare variant is enriched between two populations, since rare variants are not encountered frequently (by definition) across individuals.

For rare variant analyses, it is common to consider the entire gene as a singular unit, and consider all (or any) rare variants in a given gene as affecting gene function. These statistical tests are called "gene burden" tests.

Gene burden tests can be roughly placed into three categories. Collapsing tests collect all of the observed rare variants in a single gene in a given sample, and generates a binary outcome for the gene. The most popular collapsing test is the CMC test — the CMC scores genes as affected (1) if there is one or more rare variants observed in a sample, and zero otherwise (no rare variants observed) [145].

Aggregating tests treat each variant individually, often assigning weights to the predicted strength of each variant. These variants are then summed together to determine the aggregate "burden" for a given gene in each sample. The weighted sum statistic (WSS) was the first weighted aggregating test, where rarer variants were given a heavier weight [146].

Currently, the most popular test is the sequence kernel association test (SKAT) [147]. SKAT fits a logisitic regression model on the data, incorporating study-specific covariates, and the weighted covariance between all observed genotypes.

Since the majority of the rare variants in this study are singular variants, computing the genotype covariance matrix might be of only limited value. Instead, I decided to use the simplest method, the CMC test, and consider a gene to be affected if there are one or more rare variants observed in a sample. Since this is a binary outcome, I created a contingency table of affected/unaffected controls/schizophrenics, and used Fisher's exact test to test the probability that gene affectivity is independent from disease status. I also calculated the proportional difference between cases and controls. This value is positive if more schizophrenics are affected than controls, and negative if vice versa. I also ignored genes that had less than 5% gene affectivity in both groups, since this suggests a lack of data for that particular gene. Unfortunately, after multiple testing correction (by using both Bonferroni and Benjamini-Hochberg), no gene remains statistically significant [106].

Table 6.4: Genes with greatest proportional difference between schizophrenics & controls.

HUGO	CTRL proportion	SCZ proportion	Uncorrected p-value	Prop. diff.
SYNE1	0.171	0.297	0.0029	0.125
PTPRK	0.143	0.267	0.0024	0.125
PPFIA2	0.057	0.145	0.0033	0.088
LRRC4C	0.139	0.227	0.0257	0.088
DOCK3	0.163	0.250	0.0341	0.086

Figure 6.14 shows a histogram of the proportional differences of gene affectivity between the schizophrenic group and the control group. For example, a value of 0.0 means that the proportion of samples containing at least one or more deleterious rare variants in that gene is the same in the schizophrenic and control groups.

If we create a quantile-quantile plot (QQ-plot) of the proportional differences against a normal distribution, we notice that the data reflect a gaussian relatively well, with some possible outliers near the extreme positive end (Figure 6.15).

Table 6.4 lists the top five genes with the greatest proportional difference between schizophrenics and controls. Of particular note are genes SYNE1 and PTPRK, which appear to be outliers when looking at the QQ-plot, with a proportional difference of 0.125.

SYNE1, also known as enaptin, or synaptic nuclear envelope protein 1, is a nuclear envelope protein, and has been shown to interact with *DISC1*, a gene known to be implicated in schizophrenia [148, 149]. SNPs in SYNE1 have been previously associated with schizophrenia, bipolar disorder, and recurrent major

depression [150, 151].

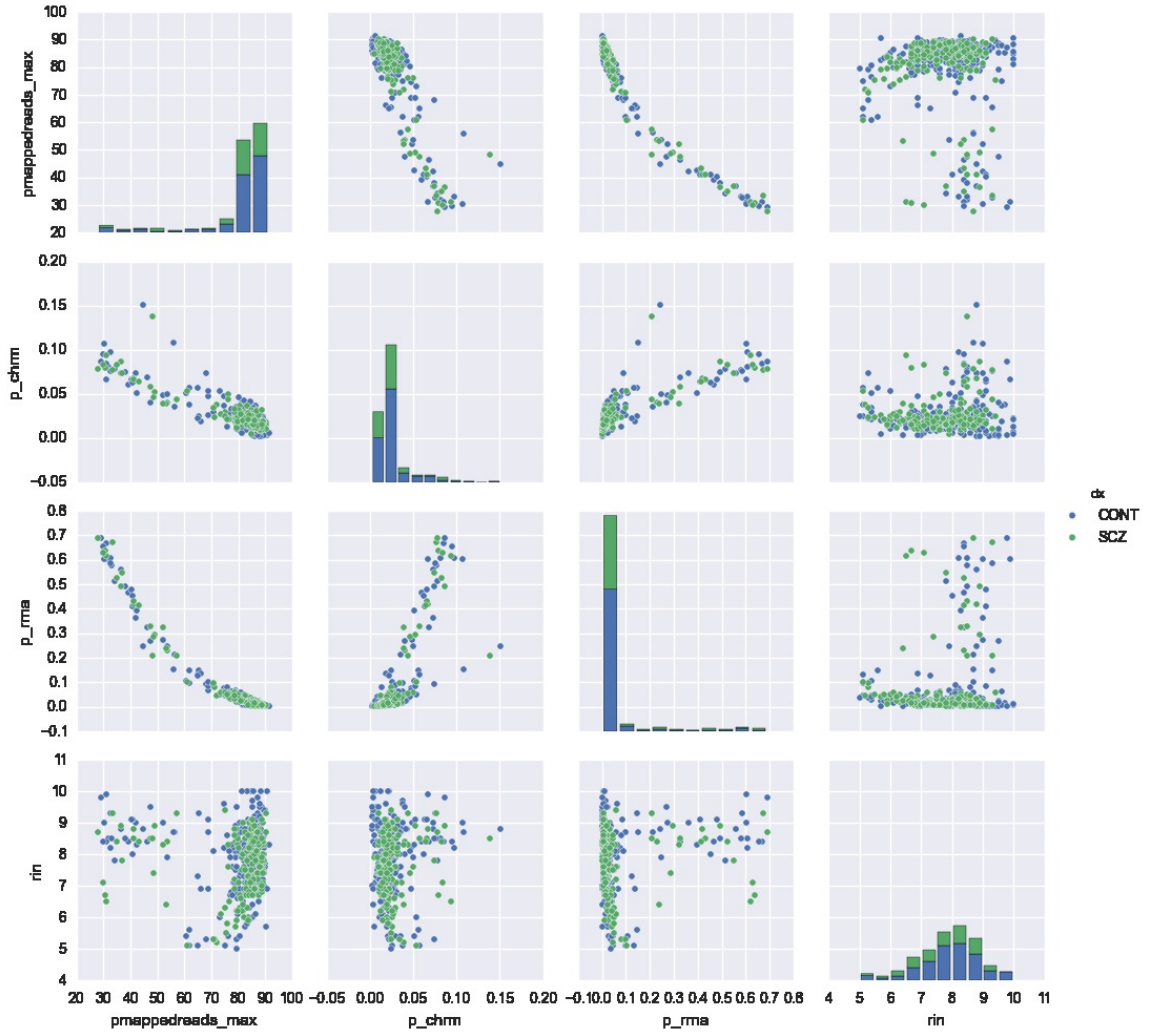

Figure 6.1: Scatterplots of RiboZero Gold RNA-seq QC metrics. Metrics plotted: Percent mapped reads, proportion of reads mapping to chrM, proportion of ribosomal RNA, RNA integrity number.

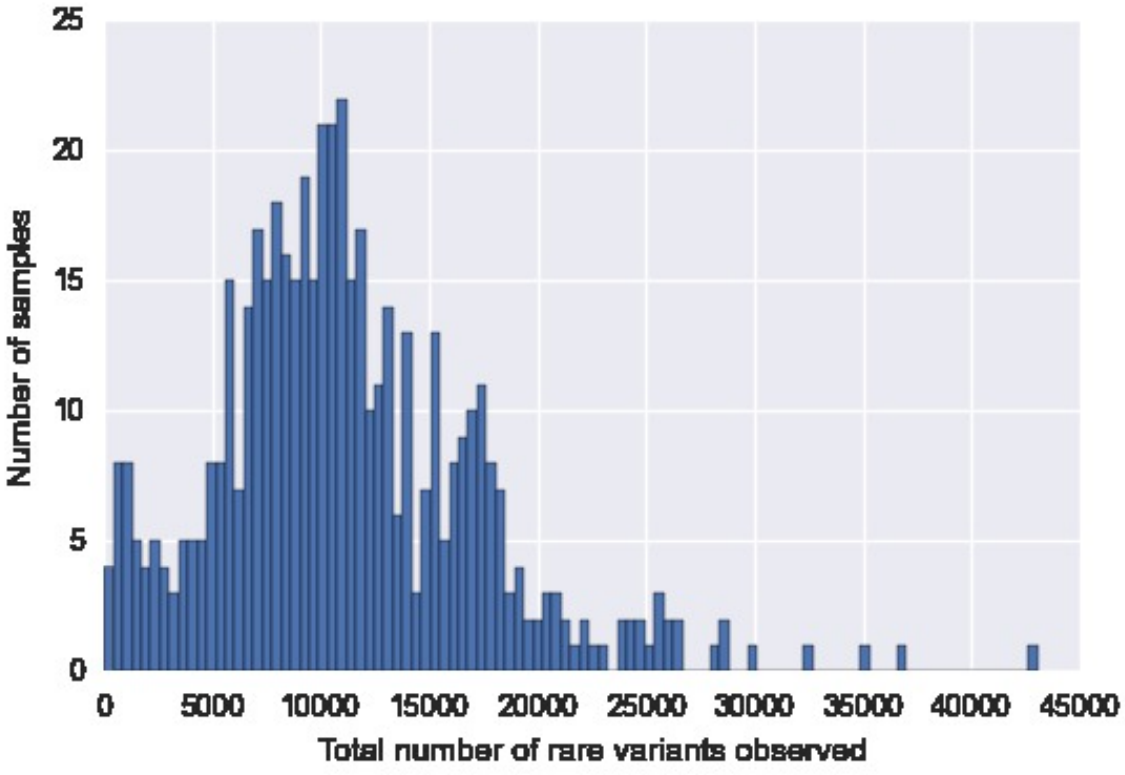

Figure 6.2: Histogram of the number of variants detected per sample.

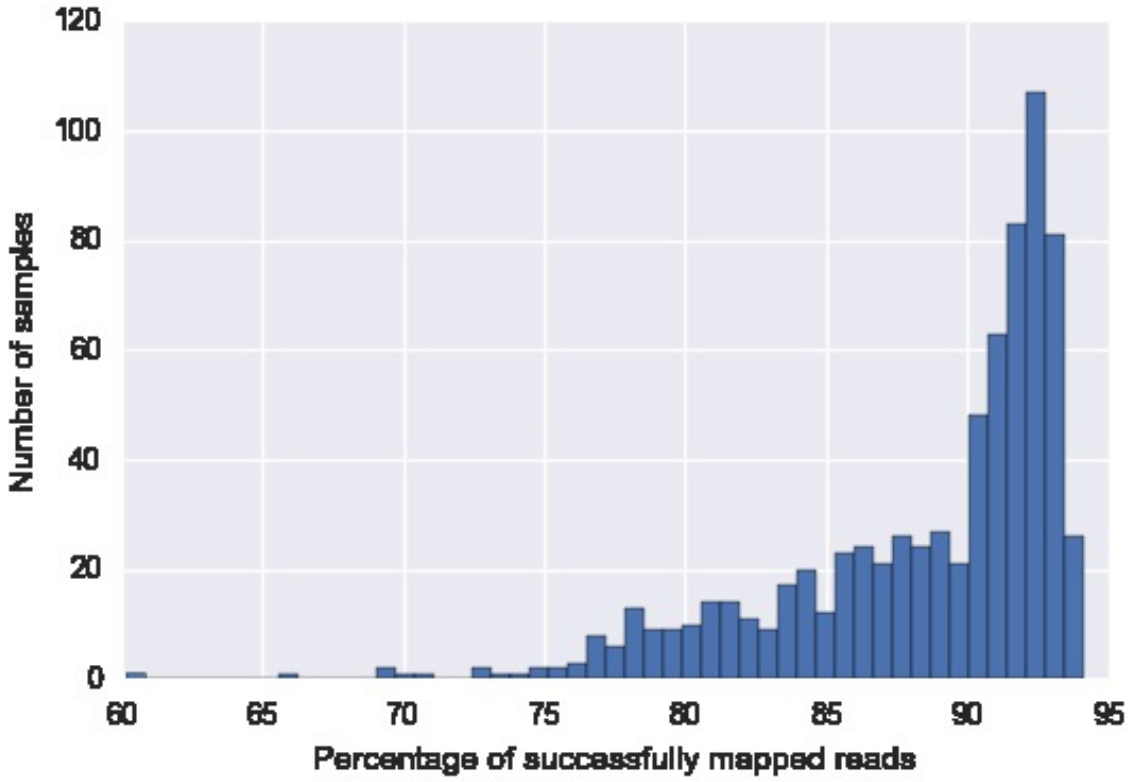

Figure 6.3: Histogram of the percentage of reads successfully mapped in the Poly-A RNA-seq data.

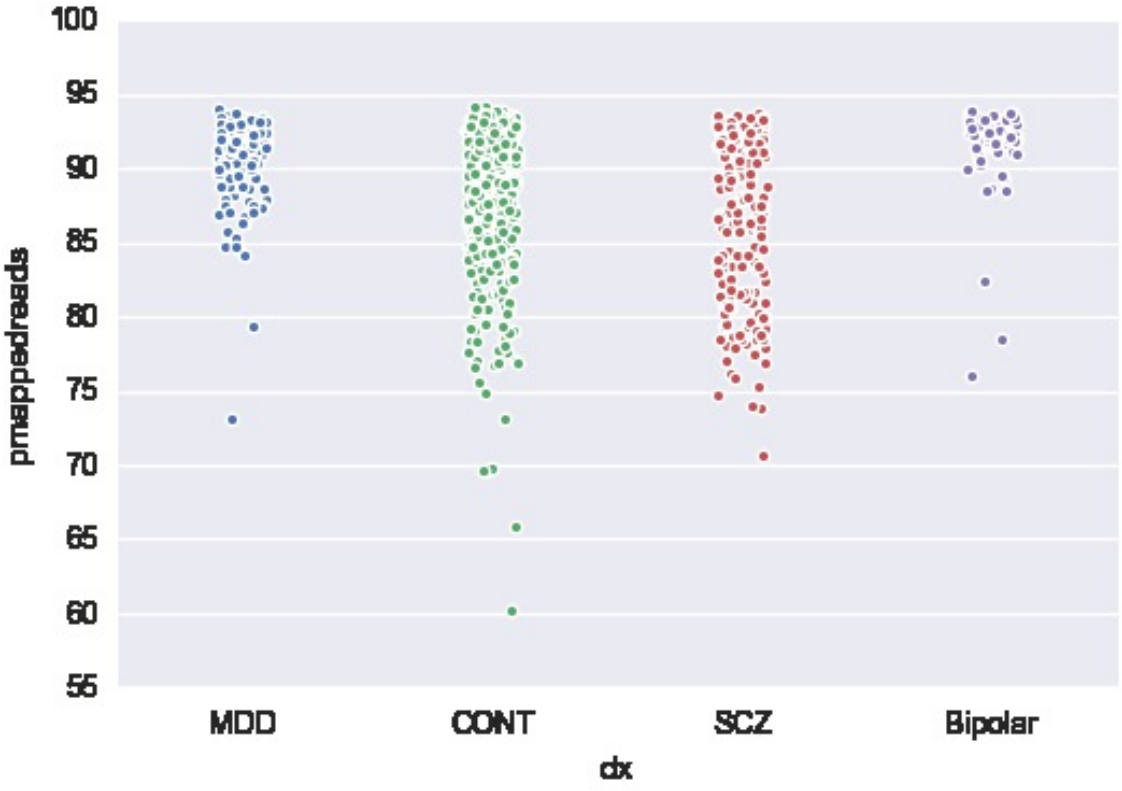

Figure 6.4: Strip-plot of mapped read percentage of Phase1 Poly-A RNA-seq data.

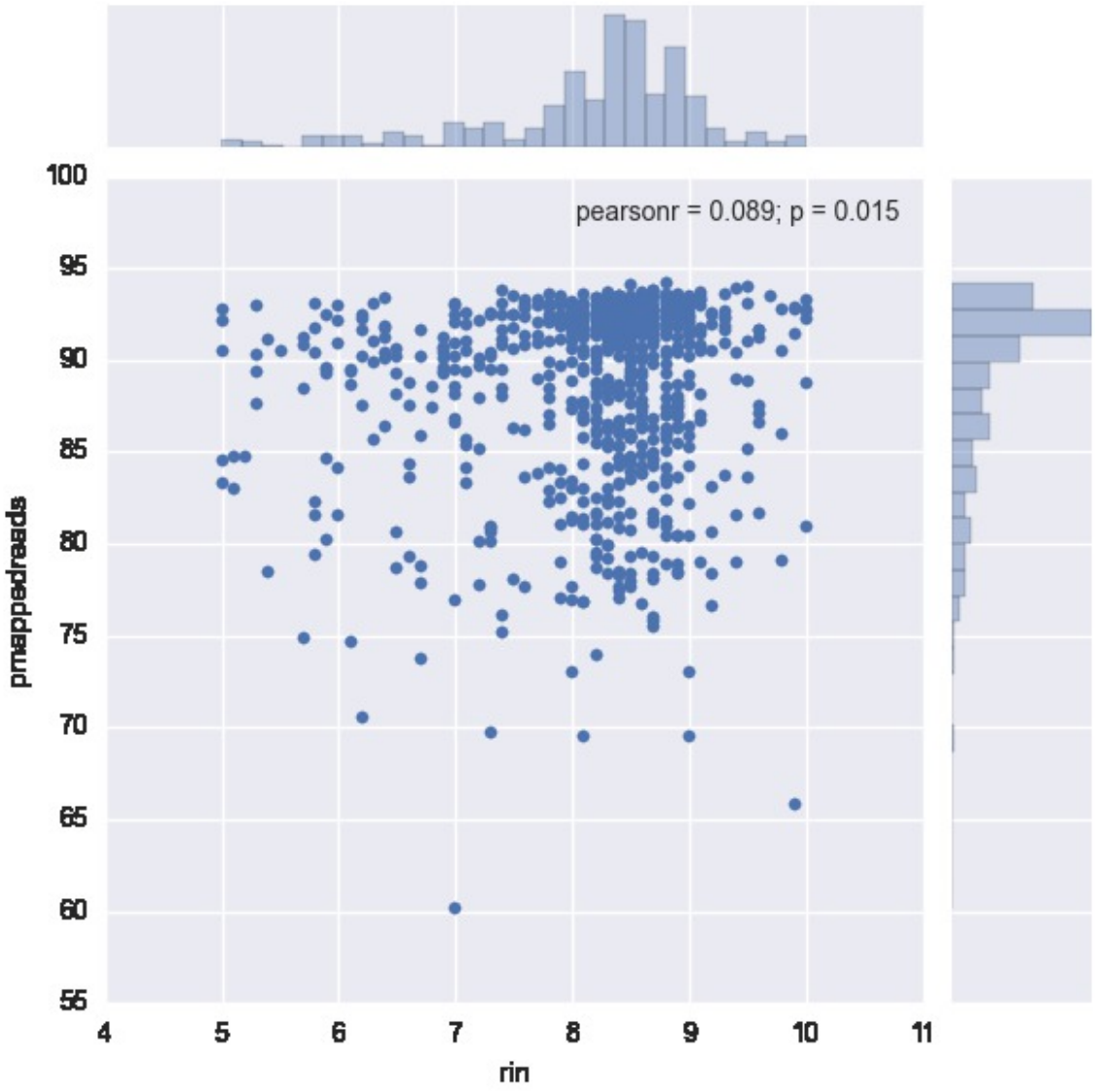

Figure 6.5: Joint density/scatter-plot of the proportion of mapped reads against the RIN.

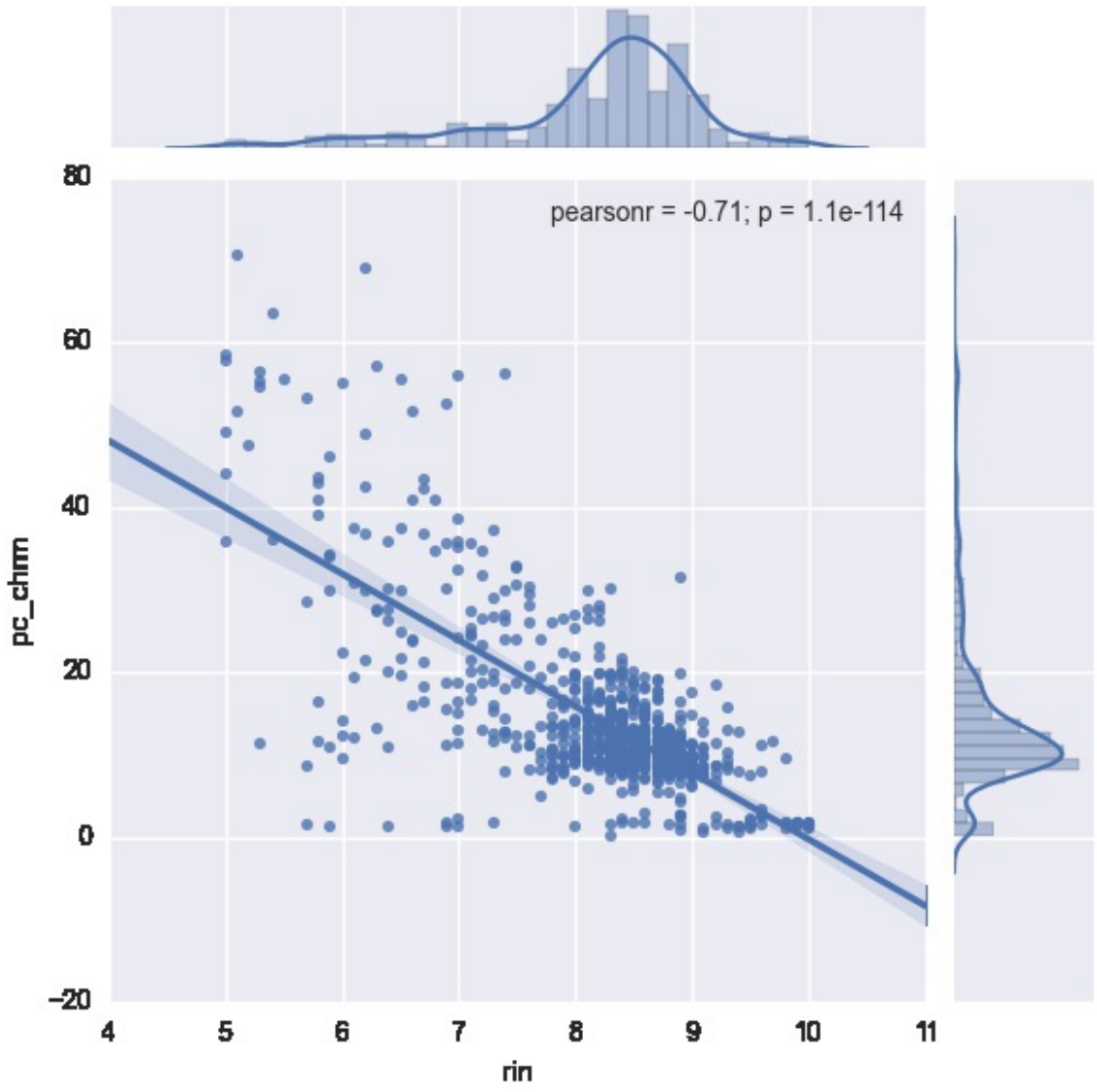

Figure 6.6: Joint density/scatter-plot of the proportion of reads mapping to chrM against the RIN.

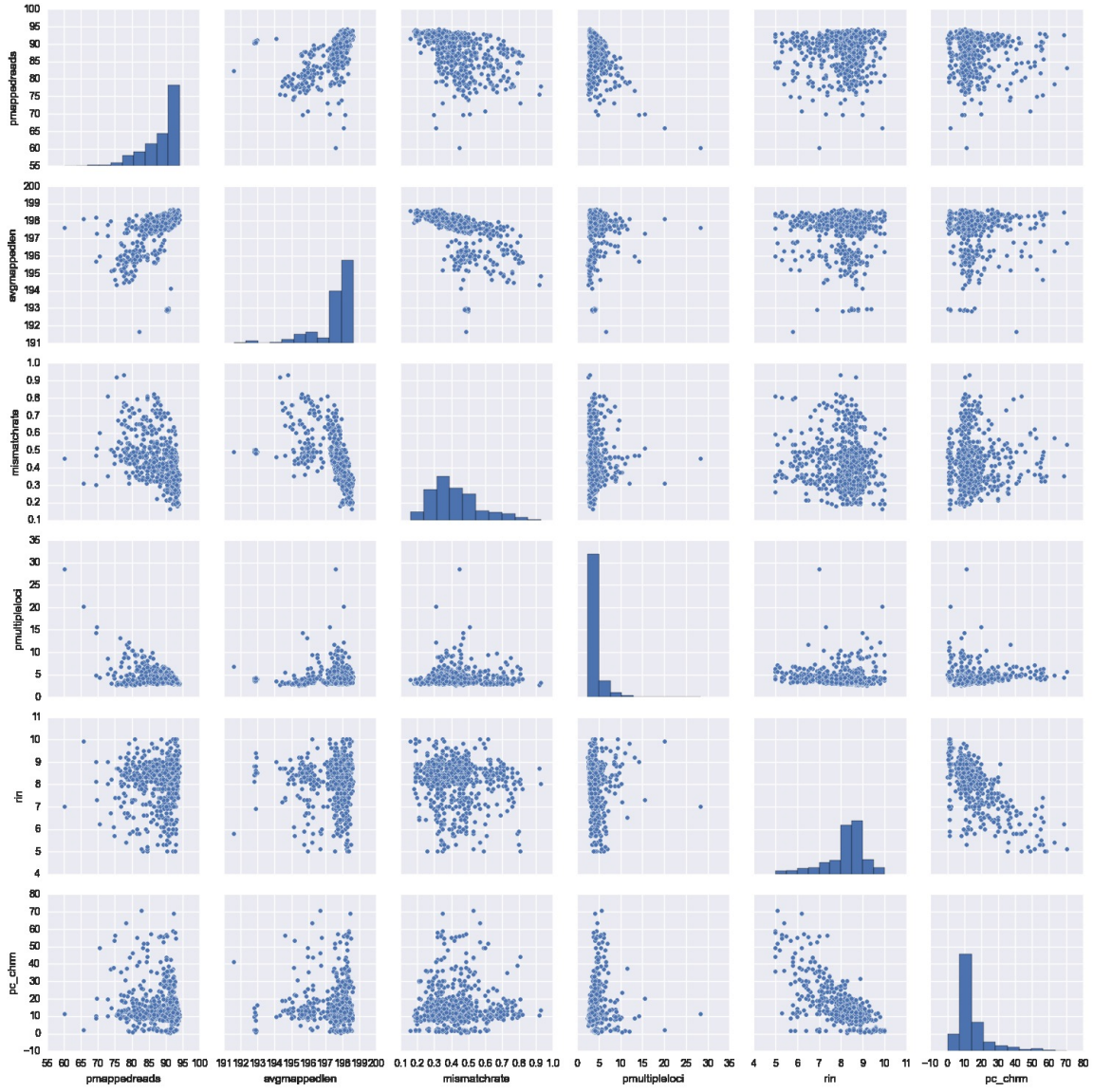

Figure 6.7: Pairwise scatterplots of five different QC metrics for Phase1 PolyA RNA-seq data.

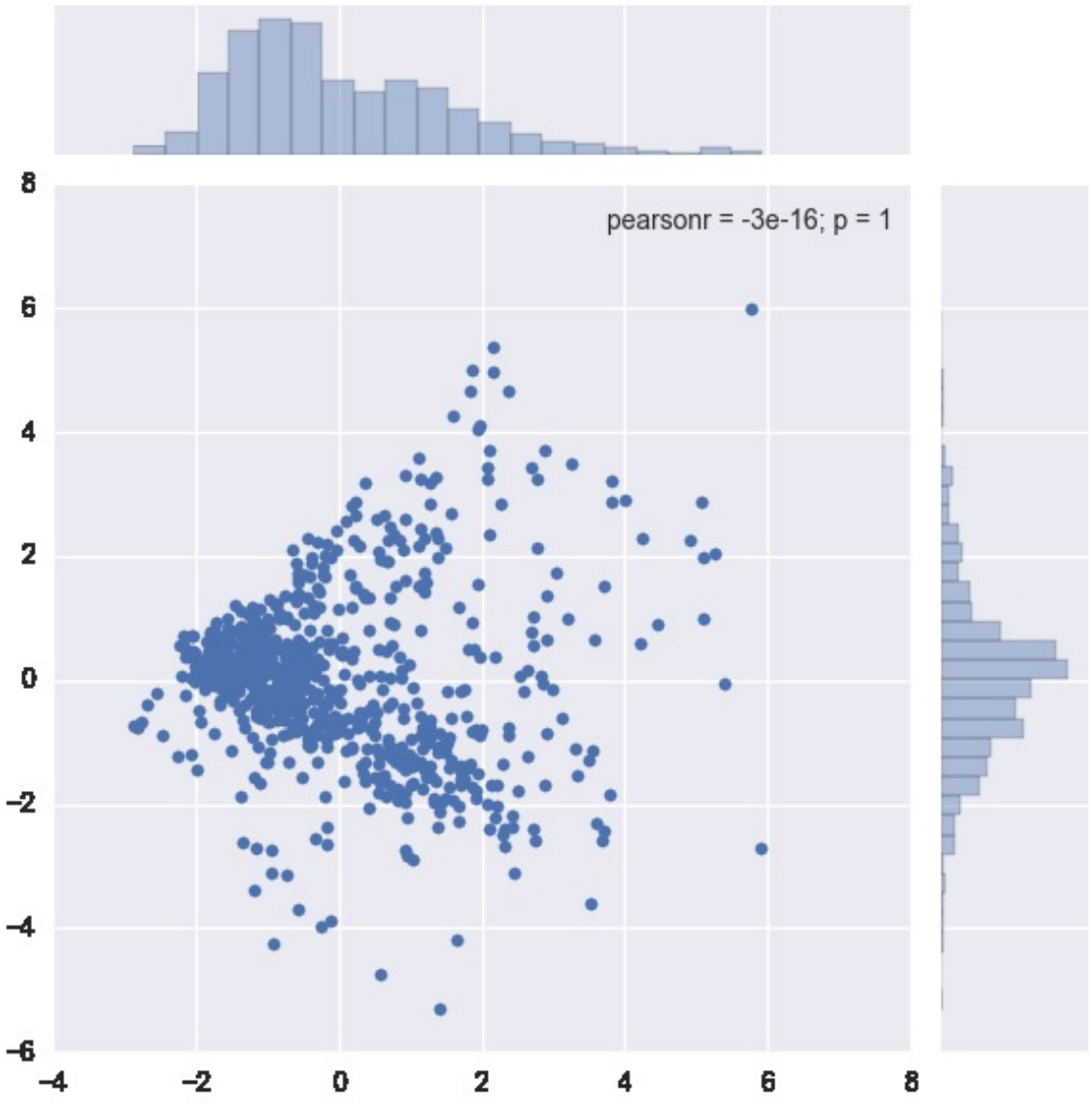

Figure 6.8: PCA of the first two components of five different QC metrics for Phase1 PolyA RNA-seq data.

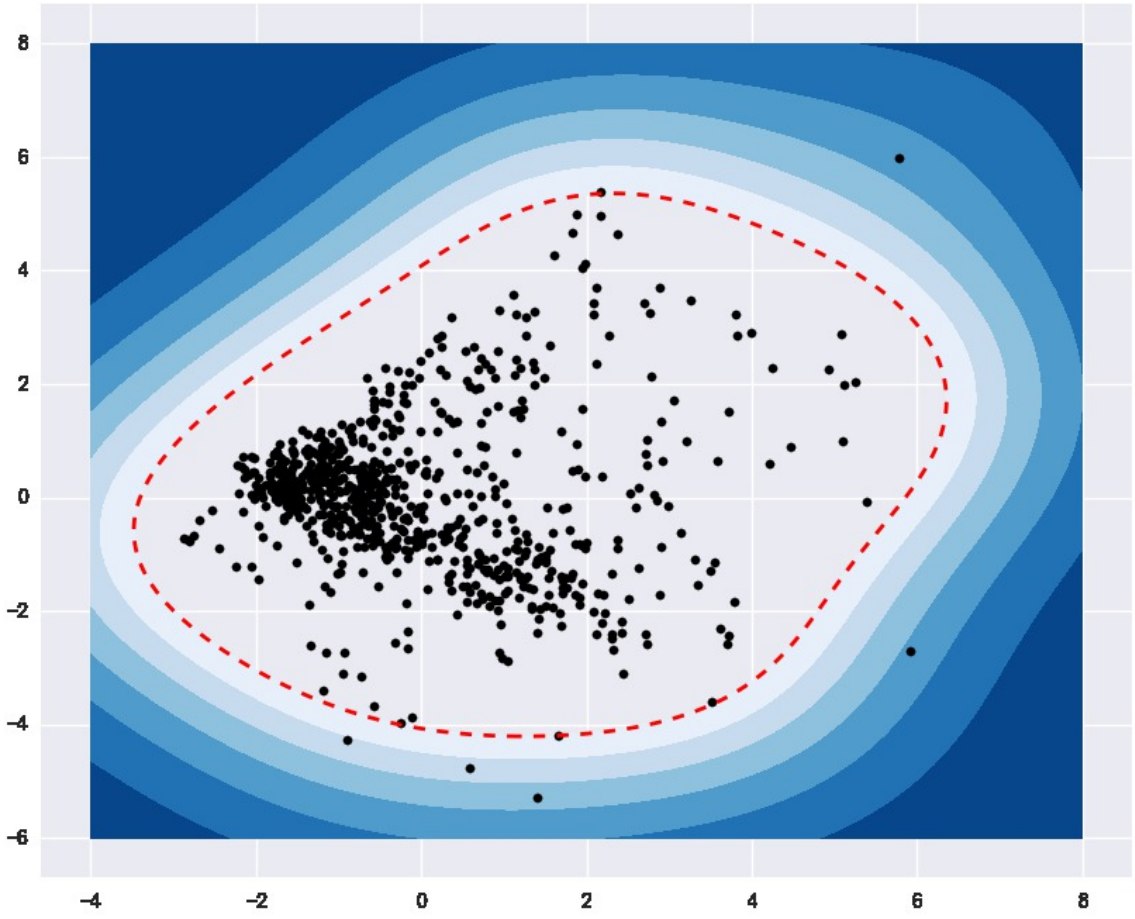

Figure 6.9: One-class SVM heatmap of the first two PCA components of the Phase1 PolyA RNA-seq QC metrics.

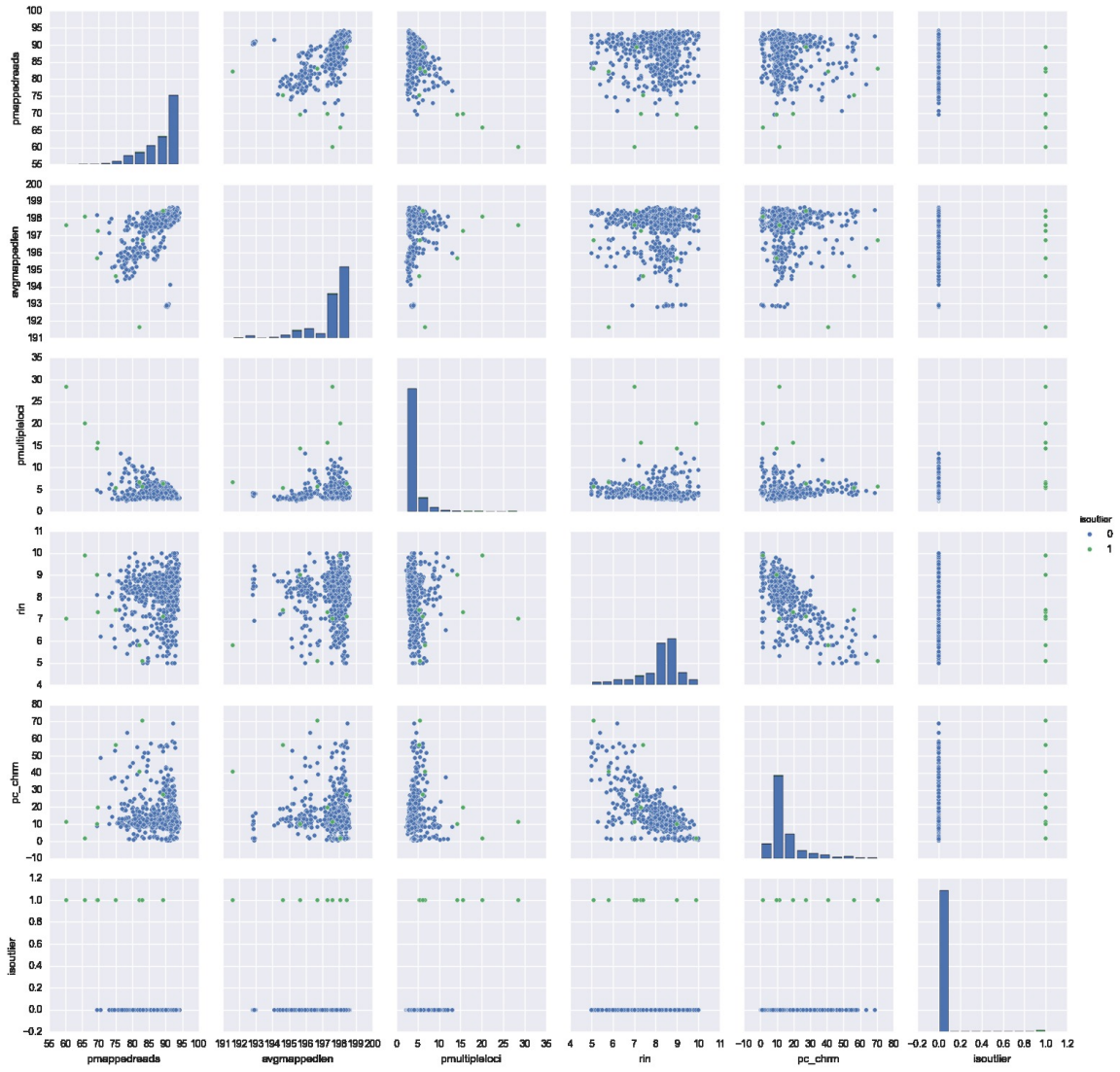

Figure 6.10: Pairwise correlation plots of 5 QC metrics of the Phase 1 PolyA RNA-seq QC metrics, colored by outlier membership.

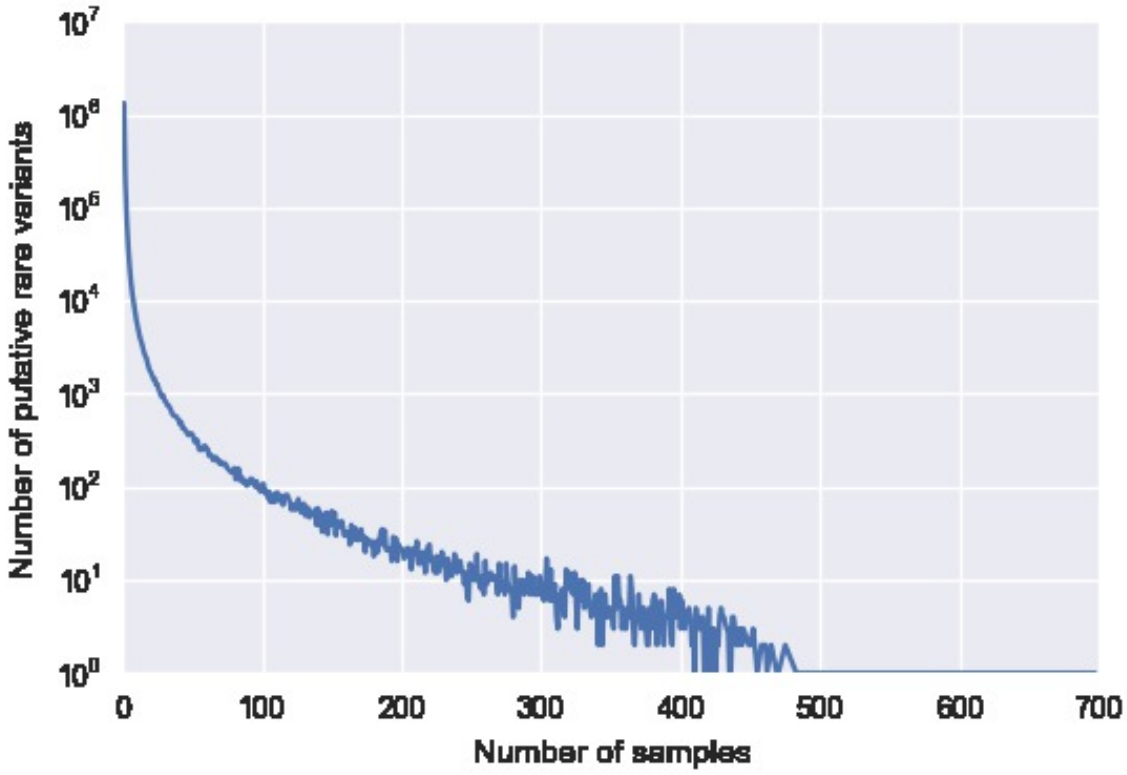

Figure 6.11: Allele frequency histogram of RiboZero RNA-seq putative rare variants. A histogram of the number of putative rare variants as a function of the number of samples containing that variant. For example, there are approximately 10 putative rare variants that are detected in 300 samples.

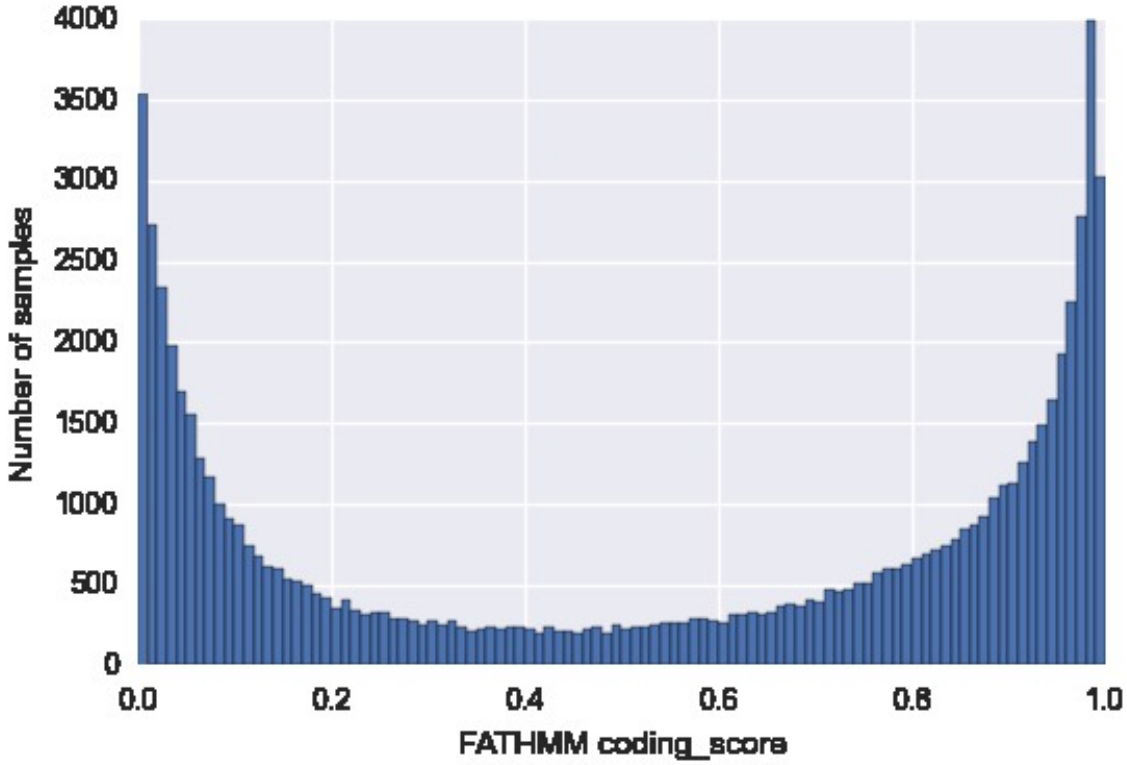

Figure 6.12: Histogram of FATHMM-MKL scores for rare exonic SNVs.

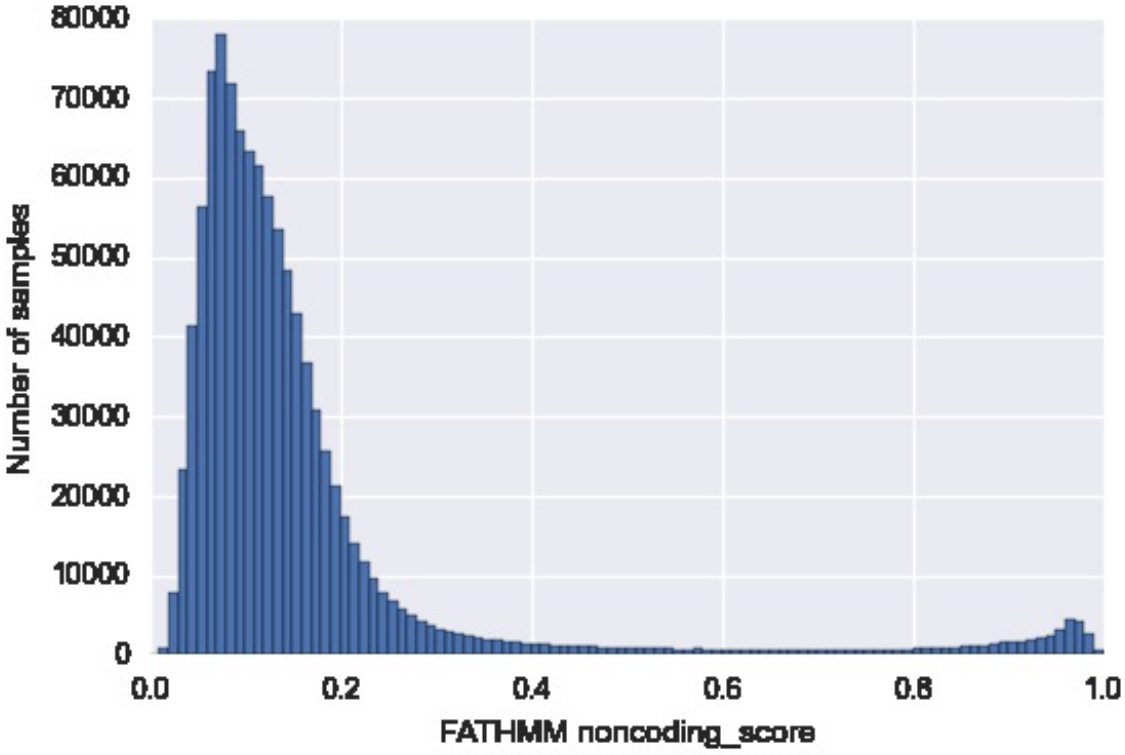

Figure 6.13: Histogram of FATHMM-MKL scores for rare intronic SNVs.

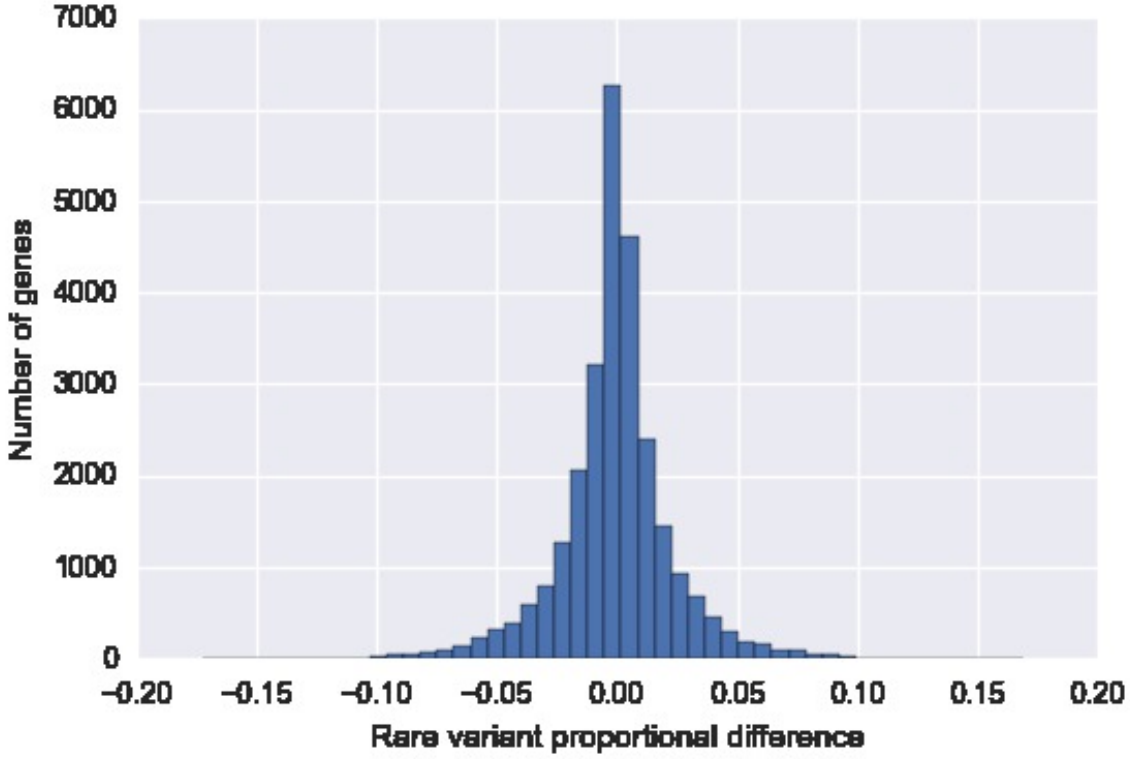

Figure 6.14: Histogram of gene affectivity differences between cases & controls.

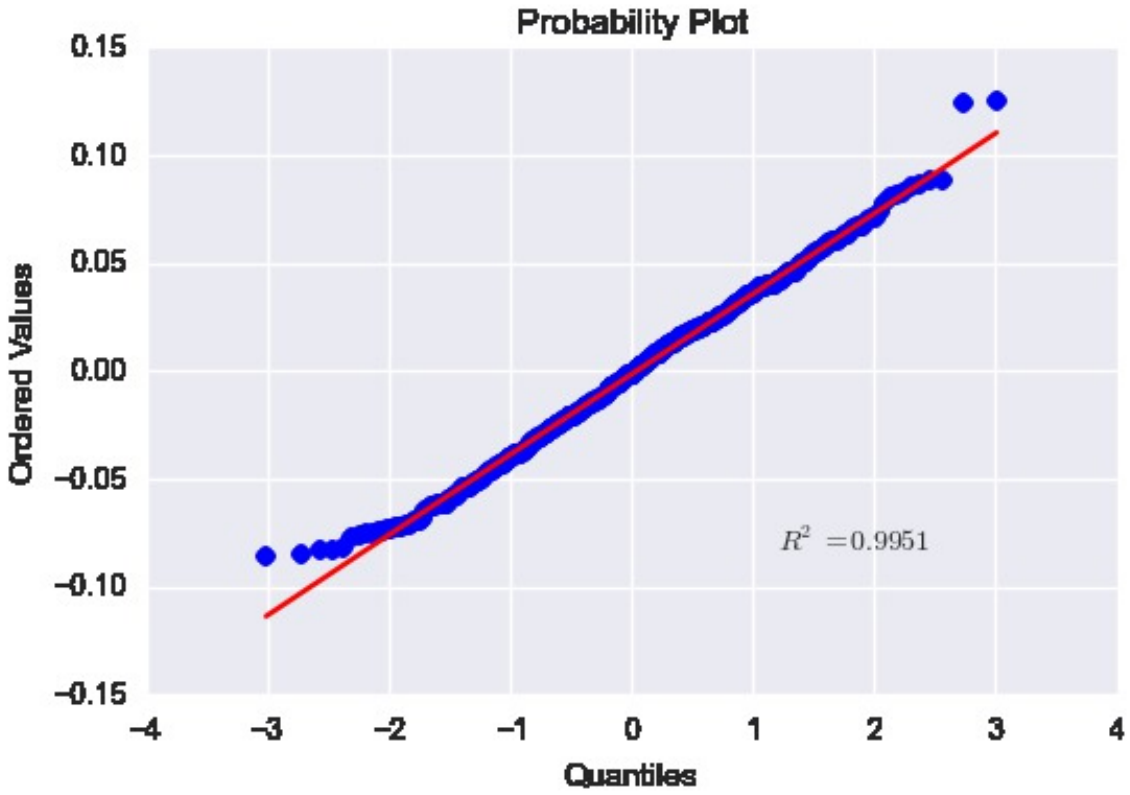

Figure 6.15: QQ-plot of gene proportional differences against a normal distribution.

Chapter 7

Analysis of short tandem repeats in schizophrenia

Variable number tandem repeats (VNTRs) are sometimes overlooked when studying the etiology of human disease. Of the different types of VNTRs, perhaps the most commonly studied type are the trinucleotide repeats in coding regions of the genome, the most well-known instance being the `CAG` repeat expansion in the huntingtin (*HTT*) gene leading to Huntington's disease [152]. A variety of trinucleotide repeats have been implicated in neurodegenerative disorders, including spinocerebellar ataxia, fragile X syndrome, and myotonic dystrophy.

Recent improvements in next-gen sequencing (NGS) technologies have made it possible for short tandem repeats to be interrogated in a genome-wide fashion [15]. With longer read lengths and highly accurate deep sequencing, STRs of

total length fewer than 85bp are eligible for detection.

Although long VNTRs (including most protein-coding trinucleotide repeat expansions) are out-of-reach using current NGS methods, STRs remain as an interesting, and possibly promising, field of study. Of particular personal interest are intronic STRs. Intronic STRs have been previously shown to alter gene function, and have been implicated in human disease. The following lists a few examples of functional intronic STRs:

- Intronic `CA` STRs near the 5′ splice-site have been shown to modulate splicing [153].
- An intronic `CA` repeat in *EGFR* ranging between 14 to 21 dinucleotide pairs has been shown to influence transcription termination closely downstream of the STR site [154].
- Perturbations of the normal length of a poly-T STR in the polypyrimidine tract near the 3′ end of intron 4 of *MRE11* are associated with significant impairment of *MRE11* expression, suggesting a link to MMR deficient tumorigenesis [155].

In this chapter, I will go over a computational pipeline for detecting STRs in NGS data (and RNA-seq data in particular), and go over some of my work trying to look for STRs significantly altered between cases & controls.

7.1 Calling STRs in NGS data

Short tandem repeats were traditionally examined by either individually screening them through Sanger sequencing, or by polling the global STR content through the use of microarrays [156]. With new NGS technologies and software pipelines, it is now possible to interrogate almost all short tandem repeats in the genome with relatively high confidence, outstripping Sanger sequencing in terms of throughput and cost, and with significant advantages over the use of microarrays, since with NGS, we are able to directly infer the genotype of each allele, instead of merely having a global picture of the abundance of each STR motif.

There are currently two bioinformatics tools for calling STRs in NGS data. RepeatSeq uses a Bayesian error model on already-aligned reads containing STRs to determine the accuracy of a particular STR candidate [157]. LobSTR works with the fastq file directly, aligning reads containing repetitive regions to the reference genome by using the sequences flanking the STR. Once the reads have been aligned to candidate STR regions, the alleles are estimated using a model of the stutter noise generated through multiple PCR runs [158].

Although these tools have been evaluated on whole genome sequencing data and whole exome sequencing data, to my knowledge, they have not been applied to RNA-seq data. In spite of this, I do not expect any serious complications, since the main difference in RNA-seq data is the inclusion of reads spanning a splice junction. In the case for lobSTR, these reads would simply appear to have a giant

deletion (of the intron), and would most likely be treated as being unalignable. There is a small risk that these reads might be misaligned, which could raise a false-positive. Again, I am not too concerned about this issue, because STRs are strongly depleted in coding regions (there are only a few thousand STRs in the coding regions of the human genome, compared to over 400,000 STRs genome-wide), so the likelihood that a junction-spanning read just happens to become misaligned, and also just happens to contain a STR is vanishingly small.

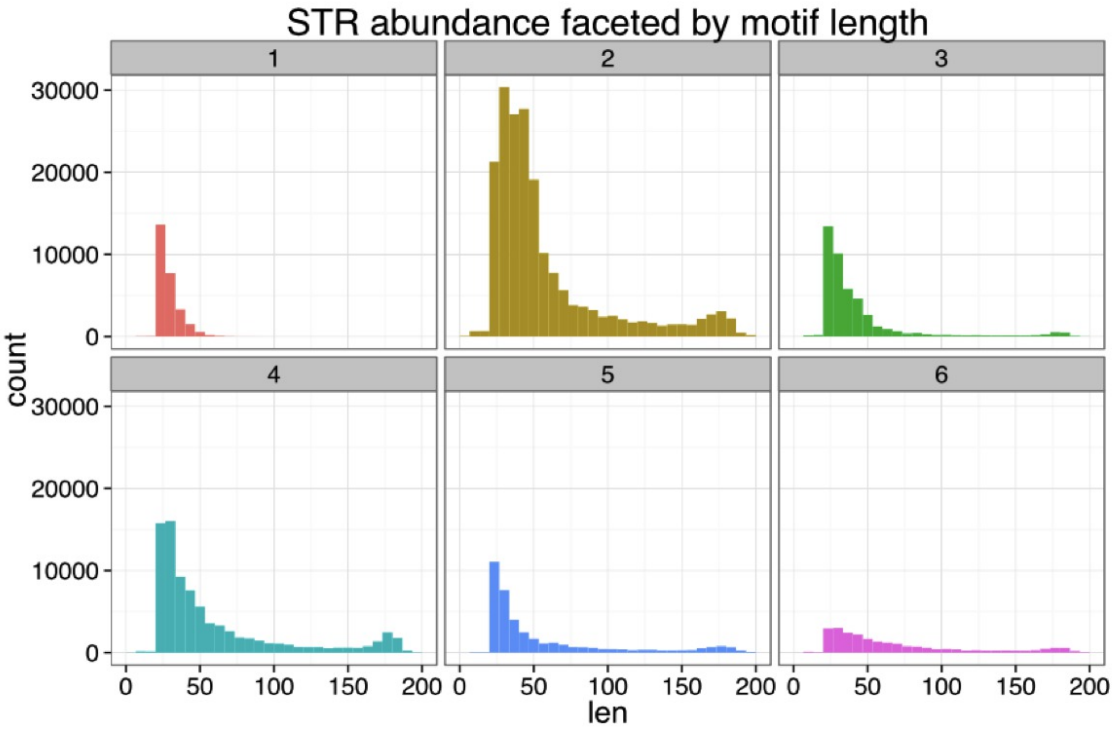

Figure 7.1: Histogram of repeatmasker STRs of different motif sizes as a function of length. STR motifs range from 1 base to 6 bases in length.

Figures 7.1 and 7.2 highlight my previous point about STR depletion in coding regions. Genome-wide, dinucleotide repeats are significantly more prevalent than

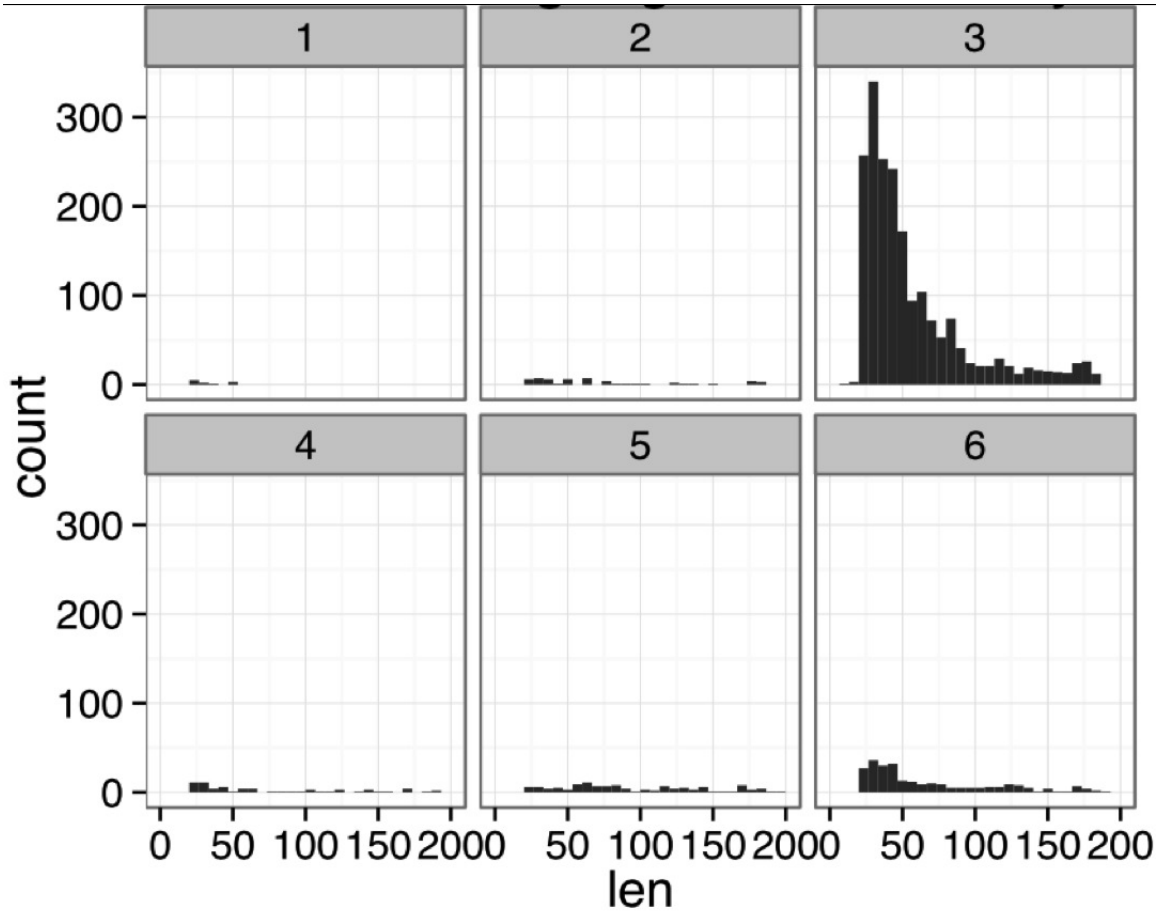

Figure 7.2: Histogram of repeatmasker STRs in coding regions of different motif sizes as a function of STR length. STR motifs range from 1 base to 6 bases in length.

other types of STRs, but in coding regions, there are practically no STRs that are not trinucleotide or hexanucleotide repeats. This is because expansions and contractions of tri- and hexa-nucleotide repeats preserve the reading frame of the transcript, but most expansions/contractions of the other STR types will result in a frameshift, most likely destroying the function of the gene. Also note that the total number of STRs in the coding region is several orders of magnitude lower than the number of STRs in noncoding regions of the genome.

For the purposes of this study, I decided to use lobSTR instead of RepeatSeq, due to the fact that lobSTR is significantly faster than RepeatSeq, and also provides tools for modeling the allelotype distribution of each STR variant.

In order to test the feasibility of calling STRs in RNA-seq data, I ran the lobSTR pipeline on a subset of the PolyA RNA-seq data. Figures 7.3 and 7.4 show the histograms of all STRs in hg19 and all detected STRs in the PolyA RNA-seq data. Figure 7.4 also shows a gradual cutoff in the number of STRs detected with total length greater than 100bp. This is due to a technical limitation of NGS: with 100bp reads, it is almost impossible to fully characterize a STR longer than 100bp, and slightly shorter STRs (between 80-95bp in length) will not have enough bases from the flanking regions for any aligner to uniquely map the read to the reference genome.

The previously provided reasons do not explain why we are able to pick up *any* STRs longer than 100bp. We are able to pick up a small proportion of $>$ 100bp STRs because we are using paired-end 2x100bp reads. If we know the insert size of the read, and the paired-end reads just happen to span the entire length of the STR (STR start at one paired-end read, STR end at the other paired-end read), then it is possible to infer the actual length of the STR, provided that the alignment is done successfully.

In order to poll STRs between schizophrenics and matched controls, I ran lobSTR with default settings on fastq files of both PolyA RNA-seq data and of

RiboZero RNA-seq data. The fastq files were stored and analyzed on a Dell Poweredge R910, configured with 1TB of DDR3 RAM, containing four ten-core Intel Xeon processors, and capable of processing 80 threads in parallel. I ran lobSTR jobs in parallel using `GNU parallel` [159]. I was particularly interested in the RiboZero RNA-seq data, because of RiboZero's ability to detect intronic regions with greater overall coverage. In the PolyA RNA-seq data, I was able to create genotype predictions of 73,025,743 STRs in 822 samples, producing an average of 88,839 STRs per sample. In the RiboZero RNA-seq data, lobSTR was able to detect 145,254,858 STRs in 500 samples, producing an average of 290,510 STRs per sample.

Although RiboZero is capable of picking up reads from intronic regions, due to the immense length of most introns, the overall coverage of most intronic STRs is a bit spotty. Figure 7.5 highlights this issue. Although there are STRs that can be detected in all 500 RiboZero RNA-seq samples, there are also a lot of STRs that can only be detected in a handful of samples. These STRs need to be removed from downstream statistical analysis.

7.2 Storage of STRs in SQL database

With such a large number of variants detected in the data (over 145 million STRs genotyped), it is essential to store the data in a structured, index-able way.

To this end, I created a python script capable of parsing and uploading VCF files into a MySQL database. In order to parse the VCF file, I used PyVCF, which parses each VCF line into a `Record` object, which allows for easy downstream parsing [160]. I also used the SQLalchemy module for Python, which provides an extremely flexible wrapper around many different flavors of SQL, allowing for easy transition between database systems such as SQLite, MySQL, and Postgres [161].

7.3 Annotation and statistical analysis of STRs

Once the STRs are stored in the database, I also annotated each detected STR using ANNOVAR to determine its biotype. Figure 7.6 shows boxplots of the number of samples that can detect a given STR, separated by biotype. In this figure, we can see that the intergenic STRs are hard to detect, highlighted by the low number of samples. Due to RNA-seq's design, we also see that, as expected, we can call STRs in exonic, UTR3, and UTR5 areas with high sensitivity. Due to the fact that most ncRNAs are lowly expressed, we also observed a relatively low sensitivity when detecting STRs in ncRNAs. It also appears that intronic STRs also fall into this "difficult to detect" category, with a median intronic STR being detectable by about 140/500 samples.

In order to try and detect possibly significant STRs between cases and controls,

for each STR, I ran a Wilcoxon rank sum test. The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a nonparametric test that calculates the likelihood that the two samples of data originate from the same population distribution [162]. Since the ranksum test is a nonparametric test, very few assumptions need to be made on the data. For each STR, we cannot make the assumption that the distribution of alleles follow a normal distribution, or that both distributions have the same variance — therefore, we cannot, in good faith, run a t-test on the data. The ranksum test asks for these assumptions on the data:

- All data points are independent and identically distributed (iid)
- Data fall on an ordinal scale (we can say data point x is greater, or smaller than, y)
- The null hypothesis states that the distributions of both populations are identical

STR allele data can satisfy these conditions, and so I conducted 329,352 ranksum tests on the data, removing 35,097 intergenic STRs. Due to the limitations of the ranksum test, and to try and mitigate the low sensitivity of intronic STR detection, I only chose to run statistical testing on STRs that were observed in at least 40 schizophrenics and 40 controls. I chose a subset of the RNA-seq data, of caucasian individuals with age greater than 16. Individuals with age less than 16 would be ambiguous in terms of diagnosis, since the onset of schizophrenia usu-

Table 7.1: Most significant STRs in RiboZero RNA-seq case/control study.

chrom	pos	biotype	hugo	motif	ref	qval
chr16	75673678	intronic	KARS	A	19	0.0025
chr15	59643060	ncRNA_intronic	MYO1E lncRNA	AAC	5.667	0.0055
chr3	172020666	intronic	FNDC3B	AC	21.5	0.0067
chr4	56309643	intronic	CLOCK, TMEM165	A	24	0.0260
chr17	26657086	intronic	IFT20	A	17	0.0530
chr1	47125640	intronic	ATPAF1, EFCAB14	AAAG	3.75	0.0697
chr8	30554930	intronic	GSR	AAAC	11.25	0.0944
chr8	26698107	intronic	ADRA1A	A	13	0.1027
chr7	123907559	ncRNA_intronic	RP5-921G16.1	A	14	0.1041
chr3	17549509	intronic	TBC1D5	A	11	0.1050
chrX	21978155	intronic	SMS	AC	7	0.1372
chr19	9755727	intronic	C19orf82	A	16	0.1601
chr3	62251572	ncRNA_intronic	PTPRG-AS1	AC	15	0.1758
chr3	113052803	intronic	WDR52	A	16	0.2216
chr19	19780721	intronic	ZNF101	A	12	0.2276

ally happens after the age of 16. The youngest individual in our dataset diagnosed with schizophrenia is 16 years of age, which is the main motivator for my choosing of this hard boundary. Caucasians were also chosen for the initial round of STR testing, as I was concerned that analyzing the samples as a whole would be confounded by racial differences that are reflected in the STR landscape.

Figure 7.7 shows that there is a significant enrichment for low p-values than one would expect by chance — for a truly null distribution, one would expect a uniform distribution of p-values. After Bonferroni correction, four STRs remain statistically significant. Bonferroni is an overtly conservative correction technique, and it is likely that there are more STRs that should be labeled as statistically significant. The most statistically significant STRs are shown in Table 7.1.

7.4 Analysis of significant genes

The most significant STR is an intronic poly-A stretch (with a cytosine anchor) inside KARS. This STR also happens to occur at the very 3′ end of a recently discovered retained intron transcript, *ENST00000565738* (Figure 7.8 shows a snapshot of the region in the UCSC genome browser). *KARS*, also known as lysyl-tRNA synthetase, primarily acts to ligate lysine to its corresponding tRNA, but has also been shown to act as a signaling molecule, inducing the primary immune response by activating monocytes [163,164]. Mutations in the aminoacyl tRNA synthetase family have been linked to neuronal pathologies, autoimmune disorders, and cancer [165]. Interestingly, *KARS* has been shown to be necessary for HIV viral assembly, interacting with *Gag* [166]. Figure 7.9 shows the distribution of alleles of the KARS STR in our study.

The second most significant STR is an AAC repeat residing in an intron of an intronic ncRNA (RP11-356M20.3), which is itself inside the first intron of MYO1E (UCSC snapshot in Figure 7.10. *MYO1E* has been shown to regulate the TLR4-triggered macrophage response [167]. Mutations in *MYO1E* have also been associated to glomerulosclerosis [168]. This is a bit of a long shot, but in a Taiwanese matched cohort study of schizophrenics and controls, after controlling for covariates, the schizophrenic hazard ratio for the development of chronic kidney disease (CKD) was 1.25 ($p < 0.05$) [169].

Some of the other genes in the top most significant list are also interest-

ing. The CLOCK gene regulates circadian rhythm, and has previously been implicated in bipolar disorder and major depressive disorder. Schizophrenia has been previously shown to be co-morbid with sleep and circadian rhythm disruption [170–172]. IFT20 codes for a intraflagellar transport protein, and evidence suggests *IFT20* to regulate the immune synapse assembly in T cells. *IFT20* has also been shown to interact with *DISC1*, a known schizophrenia gene [149].

Perusal of the top most significant genes in this analysis all have some link to the immune response, which I find interesting, as there is a previously established link between schizophrenia and the immune response. Müller *et al.* revealed that schizophrenics suffer from impaired monocyte activation, with an abnormally high level of *TLR3* and *TLR4* receptors [173]. Several other studies also suggest impaired monocyte function in schizophrenics [174–176]. Macrophages have also been shown to follow a circadian cycle [177].

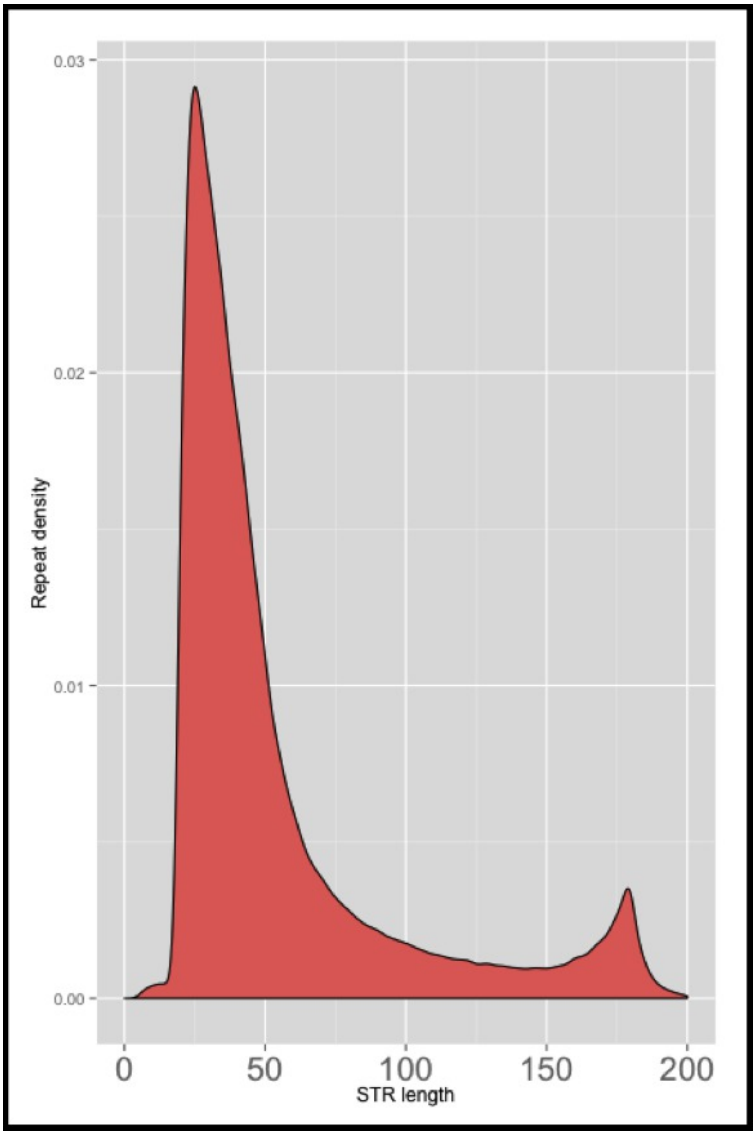

Figure 7.3: Histogram of all repeatmasker STRs in hg19 as a function of STR length.

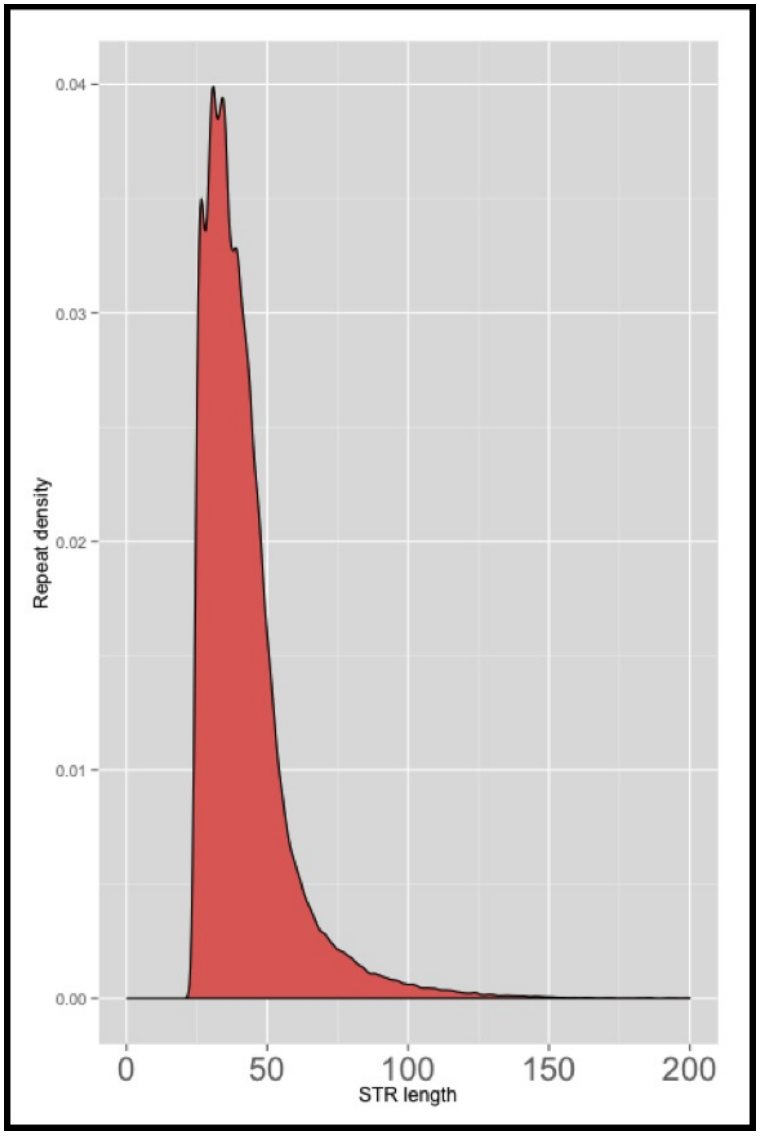

Figure 7.4: Histogram of all detected STRs in a subset of PolyA RNA-seq data as a function of STR length.

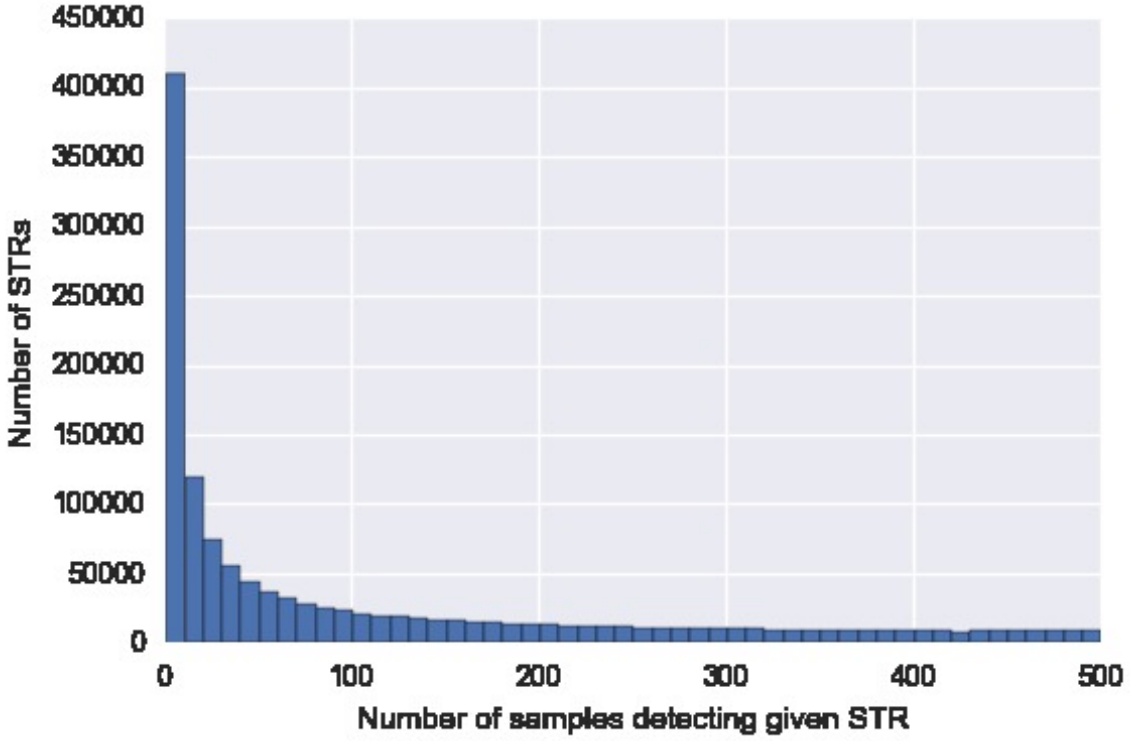

Figure 7.5: Histogram of the number of RiboZero RNA-seq samples detecting a given STR.

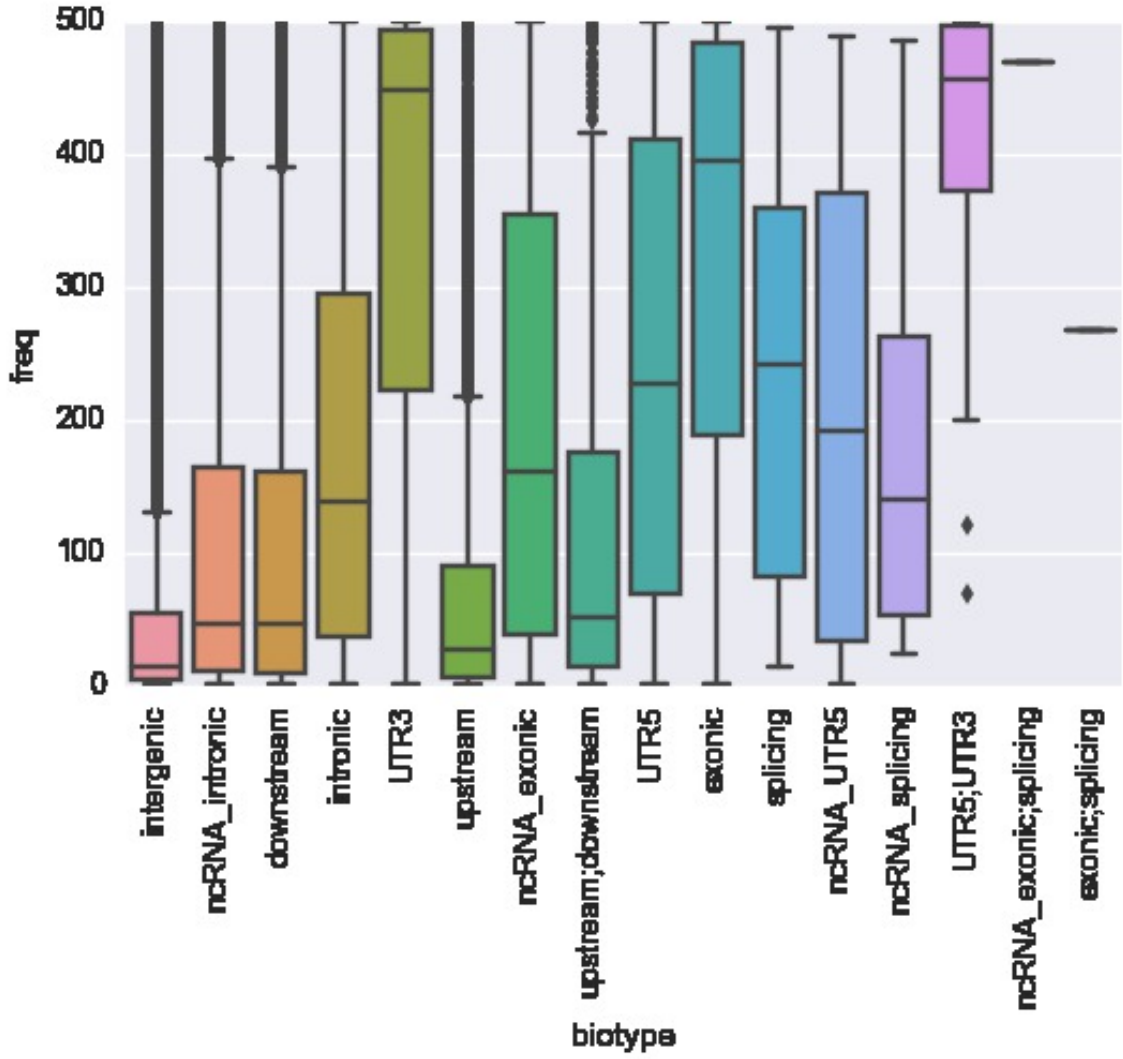

Figure 7.6: Boxplots of the number of samples that detect a given STR, separated by biotype.

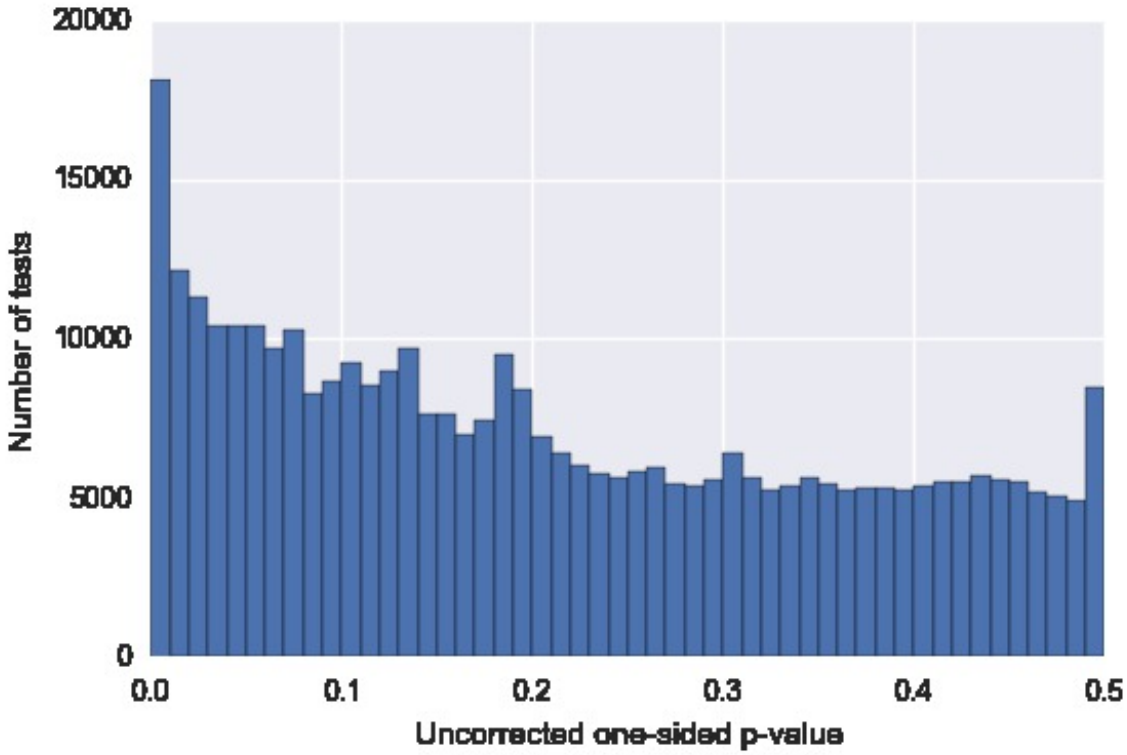

Figure 7.7: Distribution of uncorrected p-values comparing STRs of cases vs. controls.

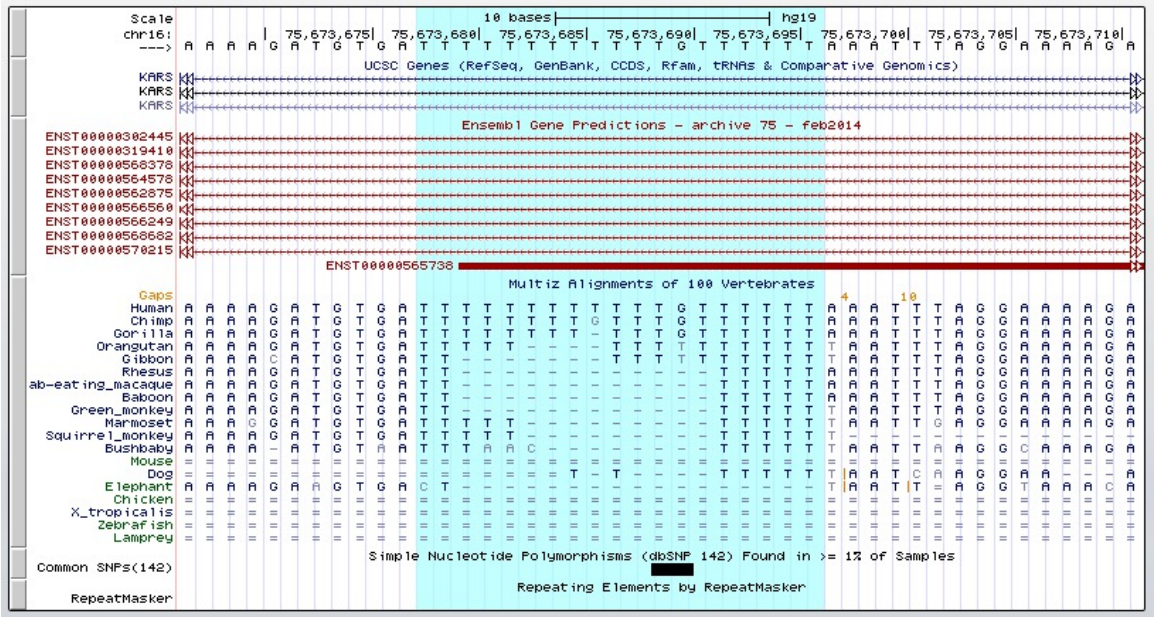

Figure 7.8: UCSC genome browser snapshot of *KARS*. Statistically significant STR is highlighted in blue.

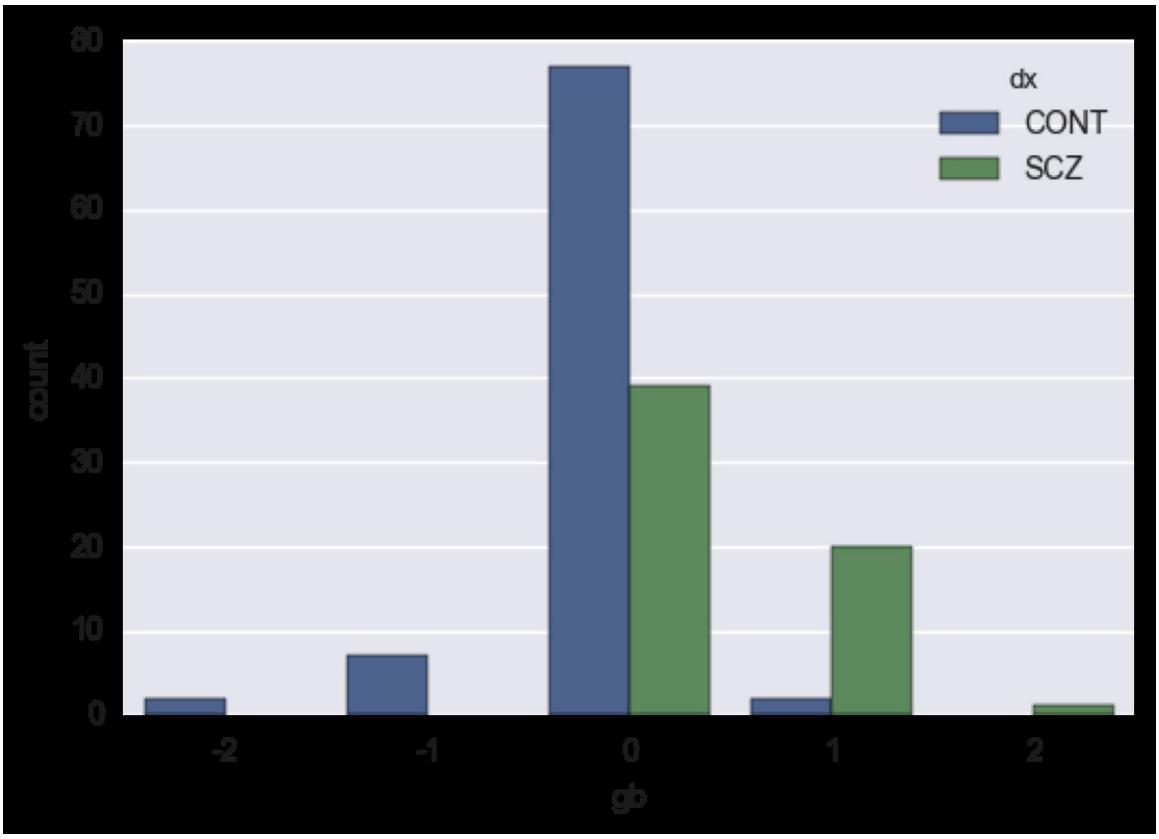

Figure 7.9: Distribution of alleles in significant *KARS* STR. The x axis measures the base length difference from hg19 reference.

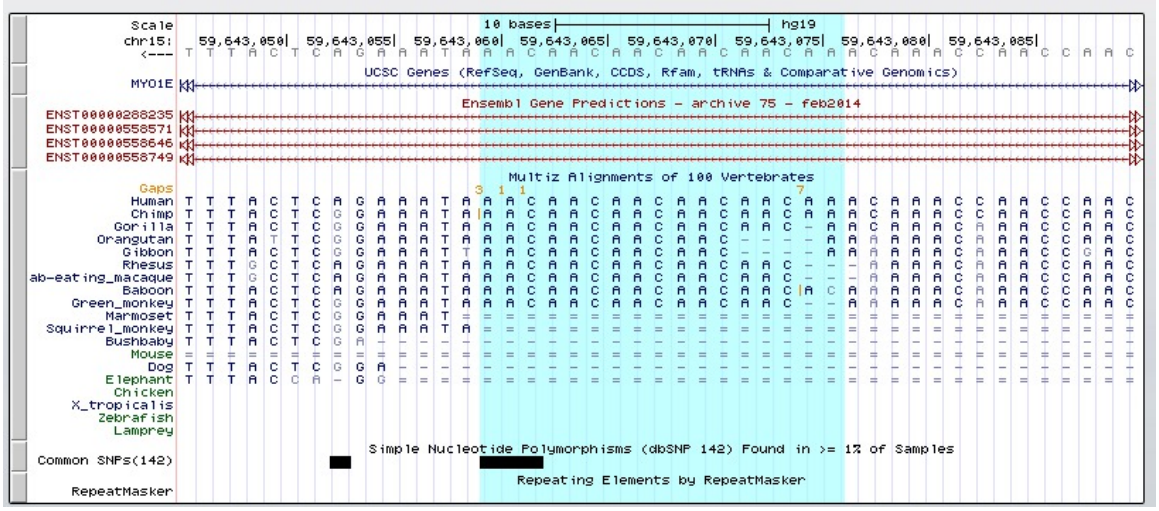

Figure 7.10: UCSC genome browser snapshot of *MYO1E*. Statistically significant STR is highlighted in blue.

Chapter 8

Future work

8.1 Refinement of rare variant calling pipeline

The three-pass rare variant calling pipeline is in a very early prototypical stage, and I have in mind various improvements and simulations that can be done that may improve the pipeline significantly.

8.1.1 Choice of aligner

For the three-pass pipeline I chose *STAR* as the initial aligner, *GSNAP* as my secondary aligner, and *BLAT* as my final aligner. It may be worthwhile to run a permutation study, using the many different RNA-seq aligners available today. I chose these aligners for my initial build of the pipeline for the following reasons:

- *STAR* is two orders of magnitude faster than any other split-read aligner. I lacked the time and computational resources to use a different aligner for the first pass of calling rare variants. Fortunately, *STAR* turns out to perform quite well, with sensitivity and specificity similar to that of TopHat.

- *GSNAP* has the best alignment specificity. Although slower than most RNA-seq aligners, *GSNAP* does a good job reducing the number of false positives.

- *BLAT* has great sensitivity. Since I wanted to make sure candidate variant regions are uniquely mappable, I used *BLAT* to discover as many alignment alternatives as possible.

8.1.2 Choice of variant caller

I chose *Platypus* as my variant caller, since it was up to 10 times faster than the next fastest candidate, *freebayes*. Ideally, one would most likely want to choose a high-sensitivity variant caller for the "first pass," then go for a variant caller with great specificity (and sensitivity, if possible) on subsequent passes. *Platypus* had great specificity, but the hit to sensitivity was quite extreme, at around 50%. In future runs, I would most likely test *freebayes* first, since it looks like freebayes has good sensitivity, and then perhaps run the second pass using GATK's *HaplotypeCaller*. *HaplotypeCaller* has a reputation for having both good sensitivity and specificity, but due to speed requirements, I decided to go for Platypus instead.

8.1.3 Incorporation of variant calling into the alignment process

I think it is a mistake for variant callers to conflate rare variants with common variants. I imagine that the primary aim for most users of NGS variant-calling tools are using them to exclusively find rare variants, not common variants. If somebody is only interested in common variants, then they would save a lot of money using a SNP chip, instead of NGS.

While I was writing this dissertation, attempting to figure out summary statistics for common and rare variants, I thought of an idea that may streamline the variant-calling process. Earlier in this thesis, I wrote that the average individual has about 3.4 million SNPs, and about 300,000 rare variants. That means that the typical variant-calling pipeline spends more than 90% of its time trying to characterize *common* variants, even though over 90% of the users are probably only interested in the *rare* variants. Since, by definition, these common variants are encountered frequently in sequencing studies, perhaps we can "anticipate" the arrival of these common variants and include them as part of the reference genome, when building indexes for alignment software. In other words, in addition to the typical chromosomal fasta entries in the reference, we could add in 4.3 million "pseudochromosomes" of SNPs (and their flanking regions) into the human reference, and then align reads to this combined reference. By doing this, we

can achieve two things. We quickly dispatch the discovery of common variants, hopefully reducing the workload of the downstream variant caller by an order of magnitude. We also can improve the specificity of calling these common variants, since misalignments will be addressed as part of the index-building process.

Another advantage of this step is that it is variant-caller agnostic; this can be a stand-alone tool that will dispatch the known common variants in a given fastq file, and then the user can choose a favorite variant caller tool to process the subsequent BAM file for rare variants, with the satisfaction that the variant-calling tool might perform up to 10x faster. Of course, the alignment step will be significantly slower, but since the alignment step is already extremely fast (especially when using STAR), I think the loss in alignment speed will be more than made up by the gains in variant calling speed.

8.2 Analysis of rare variants in schizophrenia

I called the rare variants in the case/control schizophrenia RNA-seq dataset using my prototype three-pass pipeline. In the previous section, I have listed several improvements for the pipeline, and I believe improvements to that pipeline may make significant improvements to this analysis as well. Since I took a 50% hit to sensitivity by using *Platypus*, changing the variant caller to *freebayes* or *Hap-*

lotypeCaller may significantly improve statistical power.

8.2.1 Use of other variant annotation/interpretation tools

For my initial analysis, I used FATHMM-MKL to select for putative functional variants. One possible weakness of using FATHMM-MKL is that it is trained on mainly features originating from ENCODE. It might be worthwhile to incorporate other variant annotation tools. One tool that comes to mind would be a recent tool that uses a deep neural network to discover many intronic variants that affect splicing [141]. It might also be worthwhile to incorporate other features, such as those regions found to interact with *MECP2* or *TDP-43*.

8.2.2 Using a more sophisticated gene model

Instead of using CMC, which is the simplest gene burden method, it might be worthwhile to use an additive method that perhaps incorporates the variant scores from various annotation tools. It might also make sense to try and incorporate positional clustering as a possible feature, in the sense that functional rare variants may tend to cluster around functional regions of a gene.

8.2.3 Incorporation of gene fusions and expression data

RNA-seq is also capable of detecting gene fusions, which tend to significantly alter the genes. Since CNVs have been implicated in schizophrenia in the past, it might be worthwhile to also determine the impact of gene fusions in schizophrenia.

Using the RNA-seq expression data would also add another dimension to the analysis. If the existence of a rare variant also corresponds to a significant alteration of alternative splicing, this might be evidence to suggest that the rare variant in question might be functional. Care must be taken when doing this analysis, though. Using the same reads to quantify alternative splicing as well as doing rare variant calling might introduce hidden confounders that affect both variables. One way to mitigate this would be to use the set of brain samples that have had both Poly-A RNA-seq and RiboZero RNA-seq. The RiboZero data would be used to call the variants, and the Poly-A data could be used to quantify alternative splicing.

8.3 Determining viral load using RNA-seq

This is probably a long-shot, but in light of some "out-there" hypotheses positing a link between early viral exposure to schizophrenia, I thought it might be worthwhile to quickly take a peek at the RNA-seq data that we have, to get an

idea if there's any sort of viral signature in the dataset. Executing this would be a pretty simple matter. All of the data has already been aligned to the human reference using TopHat. TopHat, by default, outputs a file of unmapped reads — reads that failed to map to the human genome. It would be a relatively simple matter to download all known viral genomes, index the file, and align the remaining unmapped reads to this reference. Some tweaks would be necessary, such as allowing the aligner to map to multiple viral genomes, since we don't want to throw away any reads that happen to map to multiple areas.

Once the alignment is complete, we can collect the data, calculate expression values (perhaps a metric such as TPM would be more informative), and look to see if there is any viral reads, and if so, whether any particular virus is enriched between cases and controls.

8.4 Analysis of short tandem repeats in schizophrenia

The short tandem repeat analysis has one major weakness, which is the low sensitivity for calling most intronic variants in the RiboZero data. Unfortunately, this is a weakness of the sequencing technology, and short of increasing the numer of mapped reads (which is already a pretty deep 80-100 million reads per sample), there is not much we can do about this, unless we transition to whole genome

sequencing.

The short tandem repeat analysis, as well as the rare variant analysis, suffer from a major confounder, which is expression: If the gene is not transcribed, there is no way RNA-seq can call variants in the gene. If the gene is differentially expressed between cases and controls, then this may also bias rare variant detection. One way to mitigate this would be to only select for genes/regions that have relatively medium-to-high expression, so that the detection rate minimally suffers from confounding. I tried to account for this in the data by enforcing a high coverage cutoff, but this only mitigates the danger.

8.4.1 Development of a more sophisticated STR allelotype model

For my initial STR analysis, I made practically no assumptions about the distribution of alleles of a given STR, and used the Mann-Whitney U test. Instead, I think it is worthwhile to use the thousand genomes data to get a prior distribution of alleles for a given STR. The control samples could then be used to verify the "null" distribution, and we can then get a more meaningful measure of STR variability.

8.4.2 Validation of called STRs

Since RiboZero RNA-seq only provides modest intronic coverage, for a given STR, there will most likely be a bunch of samples where there was insufficient coverage to determine the STR of interest. It might be worthwhile to validate these regions using vanilla sanger sequencing. Perhaps there might be a way to create a "STR-capture" technique similar to exome capture techniques.

Bibliography

[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big Data: Astronomical or Genomical?" *PLOS Biology*, vol. 13, no. 7, p. e1002195, 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.1002195

[2] WHO, *Global status report on noncommunicable diseases 2014*. World Health Organization, 2014. [Online]. Available: http://www.who.int/nmh/publications/ncd-status-report-2014/en/

[3] P. Gringras and W. Chen, "Mechanisms for differences in monozygous twins," *Early Human Development*, vol. 64, no. 2, pp. 105–117, 2001.

[4] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, a. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, a. Sheridan, C. Sougnez, N. Stange-

Thomann, N. Stojanovic, a. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, a. Coulson, R. Deadman, P. Deloukas, a. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, a. Hunt, M. Jones, C. Lloyd, a. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, a. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. a. Marra, E. R. Mardis, L. a. Fulton, a. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, a. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, a. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. a. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, a. Fujiyama, M. Hattori, T. Yada, a. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, a. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, a. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, a. Madan, S. Qin, R. W. Davis, N. a. Federspiel, a. P. Abola, M. J.

Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. a. Evans, M. Athanasiou, R. Schultz, B. a. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. a. Bailey, a. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. a. Jones, S. Kasif, a. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, a. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, a. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, a. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, a. Felsenfeld, K. a. Wetterstrand, a. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, and Y. J. Chen, "Initial sequencing and analysis of the human genome." *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[5] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. a. Nickerson, and J. Shendure, "Targeted capture and massively parallel

sequencing of 12 human exomes." *Nature*, vol. 461, no. 7261, pp. 272–276, 2009. [Online]. Available: http://dx.doi.org/10.1038/nature08250

[6] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. Van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, "GENCODE: The reference human genome annotation for the ENCODE project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[7] E. Birney, J. a. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. a. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. a. Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox,

M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. a. Hirsch, E. a. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaöz, A. Siepel, J. Taylor, L. a. Liefer, K. a. Wetterstrand, P. J. Good, E. a. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. a.

Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. a. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. a. Singer, T. a. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. a. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. S. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. a. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. W. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrímsdóttir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. a. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. a. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B.

Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature*, vol. 447, no. 7146, pp. 799–816, 2007.

[8] S. R. Eddy, "Non-Coding Rna Genes and the Modern Rna World," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 919–929, 2001.

[9] ——, "Noncoding RNA genes," *Current Opinion in Genetics & Development*, vol. 9, no. 6, pp. 695–699, 1999.

[10] "Statistics about the current human GENCODE release," http://www.gencodegenes.org/stats/current.html, accessed: 2015-9-22.

[11] J. Lejeune, M. Gautier, and R. Turpin, "Study of somatic chromosomes from 9 mongoloid children," *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, vol. 248, no. 11, p. 1721, 1959.

[12] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping." *Nature reviews. Genetics*, vol. 12, no. 5, pp. 363–376, 2011. [Online]. Available: http://dx.doi.org/10.1038/nrg2958

[13] The 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 135, no. V, pp. 0–9, 2012.

[14] Y. Jiang, A. L. Turinsky, and M. Brudno, "The missing indels: an estimate

of indel variation in a human genome and analysis of factors that impede detection," *Nucleic Acids Research*, vol. 43, no. 15, pp. 7217–7228, 2015.

[15] T. F. Willems, M. Gymrek, G. Highnam, D. Mittelman, and Y. Erlich, "The landscape of human STR variation," *Genome Research*, pp. gr.177 774.114–, 2014. [Online]. Available: http://genome.cshlp.org/content/early/2014/08/18/gr.177774.114.abstract

[16] M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb, "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science (New York, N.Y.)*, vol. 299, no. 5607, pp. 682–686, 2003.

[17] S. Anderson, a. T. Bankier, B. G. Barrell, M. H. de Bruijn, a. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. a. Roe, F. Sanger, P. H. Schreier, a. J. Smith, R. Staden, and I. G. Young, "Sequence and organization of the human mitochondrial genome." *Nature*, vol. 290, no. 5806, pp. 457–465, 1981. [Online]. Available: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=7219534&retmode=ref&cmd=prlinks$\delimiter"026E30F$npapers2://publication/doi/10.3109/10717544.2014.923064

[18] G. Gibson, "Rare and common variants: twenty arguments." *Nature*

reviews. Genetics, vol. 13, no. 2, pp. 135–45, Feb. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22251874

[19] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1193–1198, 2012. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1119675109

[20] T. a. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. a. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, a. Chakravarti, J. H. Cho, a. E. Guttmacher, a. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, a. S. Whittemore, M. Boehnke, a. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. a. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19812666

[21] I. M. Johnstone and D. M. Titterington, "Statistical challenges of high-dimensional data." *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 367, no. 1906, pp. 4237–4253, 2009.

[22] J. Fan, J. Lv, and L. Qi, "Sparse High-Dimensional Models in Economics," *Annual Review of Economics*, vol. 3, no. 1, pp. 291–317, 2011.

[23] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection

methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, pp. 409–424, 2009.

[24] K. A. Wetterstrand, "Dna sequencing costs: Data from the nhgri genome sequencing program (gsp)," *Available at: www.genome.gov/sequencingcosts*, 2015.

[25] C. a. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan, "Transcriptome sequencing to detect gene fusions in cancer." *Nature*, vol. 458, no. 7234, pp. 97–101, 2009. [Online]. Available: http://dx.doi.org/10.1038/nature07638

[26] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing," *October*, vol. 320, no. 5881, pp. 1344–1349, 2008.

[27] M. N. Bainbridge, R. L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E. R. Mardis, M. D. Sadar, A. S. Siddiqui, M. a. Marra, and S. J. M. Jones, "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach." *BMC genomics*, vol. 7, p. 246, 2006.

[28] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics." *Nature reviews. Genetics*, vol. 10, no. 1, pp.

57–63, Jan. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract

[29] Z. Peng, Y. Cheng, B. C.-M. Tan, L. Kang, Z. Tian, Y. Zhu, W. Zhang, Y. Liang, X. Hu, X. Tan, J. Guo, Z. Dong, Y. Liang, L. Bao, and J. Wang, "Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome," *Nature Biotechnology*, vol. 30, no. 3, pp. 253–260, 2012. [Online]. Available: http://dx.doi.org/10.1038/nbt.2122

[30] S. Li, S. W. Tighe, C. M. Nicolet, D. Grove, S. Levy, W. Farmerie, A. Viale, C. Wright, P. a. Schweitzer, Y. Gao, D. Kim, J. Boland, B. Hicks, R. Kim, S. Chhangawala, N. Jafari, N. Raghavachari, J. Gandara, N. Garcia-Reyero, C. Hendrickson, D. Roberson, J. Rosenfeld, T. Smith, J. G. Underwood, M. Wang, P. Zumbo, D. a. Baldwin, G. S. Grills, and C. E. Mason, "Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study," *Nature Biotechnology*, vol. 32, no. 9, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25150835

[31] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. a. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. a. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson,

P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. a. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-h. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. a. Hauser, S. Hester, H. Hong, P. Hurban, S. a. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. a. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. a. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nature biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[32] S. C. Baker, S. R. Bauer, R. P. Beyer, J. D. Brenton, B. Bromley, J. Burrill,

H. Causton, M. P. Conley, R. Elespuru, M. Fero, C. Foy, J. Fuscoe, X. Gao, D. L. Gerhold, P. Gilles, F. Goodsaid, X. Guo, J. Hackett, R. D. Hockett, P. Ikonomi, R. a. Irizarry, E. S. Kawasaki, T. Kaysser-Kranich, K. Kerr, G. Kiser, W. H. Koch, K. Y. Lee, C. Liu, Z. L. Liu, A. Lucas, C. F. Manohar, G. Miyada, Z. Modrusan, H. Parkes, R. K. Puri, L. Reid, T. B. Ryder, M. Salit, R. R. Samaha, U. Scherf, T. J. Sendera, R. a. Setterquist, L. Shi, R. Shippy, J. V. Soriano, E. a. Wagar, J. a. Warrington, M. Williams, F. Wilmer, M. Wilson, P. K. Wolber, X. Wu, and R. Zadro, "The External RNA Controls Consortium: a progress report." *Nature methods*, vol. 2, no. 10, pp. 731–734, 2005.

[33] External RNA Controls Consortium, "Proposed methods for testing and selecting the ERCC external RNA controls." *BMC genomics*, vol. 6, no. June 2004, p. 150, 2005.

[34] "Fastqc," http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed: 2015-9-23.

[35] A. Dobin, C. a. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[36] S. Anders, P. T. Pyl, and W. Huber, "HTSeq A Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp.

166–169, 2014. [Online]. Available: http://biorxiv.org/content/early/2014/08/19/002824.abstract

[37] G. P. Wagner, K. Kin, and V. J. Lynch, "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples," *Theory in Biosciences*, vol. 131, no. 4, pp. 281–285, 2012.

[38] B. Li, V. Ruotti, R. M. Stewart, J. a. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinformatics*, vol. 26, no. 4, pp. 493–500, 2009.

[39] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." *BMC bioinformatics*, vol. 11, p. 94, 2010.

[40] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, N. S. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, and F. Jaffrézic, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671–683, 2013.

[41] S. Anders and W. Huber, "Differential expression analysis for sequence count data." *Genome biology*, vol. 11, no. 10, p. R106, 2010. [Online]. Available: http://genomebiology.com/2010/11/10/R106

[42] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome biology*, vol. 11, no. 3, p. R25, 2010.

[43] L. Wang, S. Wang, and W. Li, "RSeQC: Quality control of RNA-seq experiments," *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, 2012.

[44] S. Chhangawala, G. Rudy, C. E. Mason, and J. a. Rosenfeld, "The impact of read length on quantification of differentially expressed genes and splice junction detection," *Genome Biology*, vol. 16, no. 1, p. 131, 2015. [Online]. Available: http://genomebiology.com/2015/16/1/131

[45] M. Sultan, V. Amstislavskiy, T. Risch, M. Schuette, S. Dökel, M. Ralser, D. Balzereit, H. Lehrach, and M.-L. Yaspo, "Influence of RNA extraction methods and library selection schemes on RNA-seq data." *BMC genomics*, vol. 15, p. 675, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25113896

[46] D. Gaidatzis, L. Burger, M. Florescu, and M. B. Stadler, "Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation," *Nature Biotechnology*, no. April 2014, pp. 1–9, 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nbt.3269

[47] A. Belkadi, A. Bolze, Y. Itan, A. Cobat, Q. B. Vincent, A. Antipenko,

L. Shang, B. Boisson, J.-L. Casanova, and L. Abel, "Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants," *Proceedings of the National Academy of Sciences*, p. 201418631, 2015. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1418631112

[48] J. L. Weber and C. Wong, "Mutation of human short tandem repeats." *Human molecular genetics*, vol. 2, no. 8, pp. 1123–1128, 1993.

[49] L. a. Zhivotovsky, P. a. Underhill, C. Cinniolu, M. Kayser, B. Morar, T. Kivisild, R. Scozzari, F. Cruciani, G. Destro-Bisol, G. Spedini, G. K. Chambers, R. J. Herrera, K. K. Yong, D. Gresham, I. Tournev, M. W. Feldman, and L. Kalaydjieva, "The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time." *American journal of human genetics*, vol. 74, no. 1, pp. 50–61, 2004.

[50] B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, and B. Rolf, "Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat." *American journal of human genetics*, vol. 62, no. 6, pp. 1408–1415, 1998.

[51] B. Ewing and P. Green, "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities," *Genome Research*, vol. 8, pp. 186–194, 1998.

[52] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: Discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[53] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[54] "Validating generalized incremental joint variant calling with GATK HaplotypeCaller, FreeBayes, Platypus and samtools," http://bcb.io/2014/10/07/joint-calling/, accessed: 2015-10-05.

[55] "Gatk methods and algorithms," https://www.broadinstitute.org/gatk/guide/topic?name=methods, accessed: 2015-9-26.

[56] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *arXiv preprint arXiv:1207.3907*, p. 9, 2012. [Online]. Available: http://arxiv.org/abs/1207.3907

[57] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. McVean, and G. Lunter, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications." *Nature genetics*, vol. 46, no. November 2013, pp. 1–9, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25017105

[58] P. Danecek, A. Auton, G. Abecasis, C. a. Albers, E. Banks, M. a. DePristo,

R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[59] "Vcf (variant call format) version 4.0," http://www.1000genomes.org/node/101, accessed: 2015-9-26.

[60] "vcflib, a simple c++ library for parsing and manipulating vcf files, + many command-line utilities," https://github.com/ekg/vcflib, accessed: 2015-9-26.

[61] A. F. Smit, R. Hubley, and P. Green, "Repeatmasker open-3.0," http://www.repeatmasker.org/, 1996.

[62] A. Tan, G. R. Abecasis, and H. M. Kang, "Unified representation of genetic variants," *Bioinformatics*, vol. 31, no. February, pp. 2202–2204, 2015. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btv112

[63] U. Paila, B. a. Chapman, R. Kirchner, and A. R. Quinlan, "GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations," *PLoS Computational Biology*, vol. 9, no. 7, 2013.

[64] "Variant annotation tool," http://vat.gersteinlab.org/index.php, accessed: 2015-9-26.

[65] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, pp. 1–7, 2010.

[66] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[67] I. a. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations." *Nature methods*, vol. 7, no. 4, pp. 248–9, Apr. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2855889&tool=pmcentrez&rendertype=abstract

[68] H. Hu, C. D. Huff, B. Moore, S. Flygare, M. G. Reese, and M. Yandell, "VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix," *Genetic Epidemiology*, vol. 37, no. 6, pp. 622–634, 2013.

[69] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–2070, 2010.

[70] H. a. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell, "An integrative approach to predicting the functional effects of non-coding and coding sequence

variation," *Bioinformatics*, vol. 31, no. January, pp. 1536–1543, 2015. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btv009

[71] "splitnreads," https://github.com/mjafin/splitNreads, accessed: 2015-9-26.

[72] G. Ramaswami and J. B. Li, "RADAR: A rigorously annotated database of A-to-I RNA editing," *Nucleic Acids Research*, vol. 42, no. D1, pp. 109–113, 2014.

[73] A. M. Kiran, J. J. O'Mahony, K. Sanjeev, and P. V. Baranov, "Darned in 2013: Inclusion of model organisms and linking with Wikipedia," *Nucleic Acids Research*, vol. 41, no. D1, pp. 258–261, 2013.

[74] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls." *Nature biotechnology*, vol. 32, no. 3, pp. 246–51, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24531798

[75] "RNA-seq sample of na12878," http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM754335, accessed: 2015-9-26.

[76] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander,

G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature biotechnology*, vol. 29, no. 1, pp. 24–26, 2011.

[77] W. J. Kent, "BLAT The BLAST -Like Alignment Tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.

[78] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics*, vol. 26, no. 7, pp. 873–881, 2010.

[79] D. Hanahan and R. a. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[80] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2011.02.013

[81] C. J. Sherr, "Principles of Tumor Suppression," *Cell*, vol. 116, no. 2, pp. 235–246, 2004.

[82] C. M. Croce, "Oncogenes and cancer." *New England Journal of Medicine*, vol. 358, no. 5, pp. 502–511, 2008.

[83] A. G. Knudson, "Mutation and cancer: statistical study of retinoblastoma." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 68, no. 4, pp. 820–823, 1971.

[84] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes." *Nature reviews. Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[85] "Cancer gene census," http://cancer.sanger.ac.uk/census/, accessed: 2015-9-28.

[86] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. D. Jr, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[87] S. a. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell, "COSMIC: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. D1, pp. D805–D811, 2014. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1075

[88] S. Udry and N. C. Santos, "Statistical Properties of Exoplanets," *Annual Review of Astronomy and Astrophysics*, vol. 45, no. 1, pp. 397–439, 2007.

[89] "Catalogue of Somatic Mutations in Cancer," http://cancer.sanger.ac.uk/cosmic, accessed: 2015-9-26.

[90] D. Dittmer, S. Pati, G. Zambetti, S. Chu, a. K. Teresky, M. Moore, C. Finlay,

and a. J. Levine, "Gain of function mutations in p53." *Nature genetics*, vol. 4, no. 1, pp. 42–46, 1993.

[91] K. P. Olive, D. a. Tuveson, Z. C. Ruhe, B. Yin, N. a. Willis, R. T. Bronson, D. Crowley, and T. Jacks, "Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome," *Cell*, vol. 119, no. 6, pp. 847–860, 2004.

[92] M. Oren and V. Rotter, "Mutant p53 gain-of-function in cancer." *Cold Spring Harbor perspectives in biology*, vol. 2, no. 2, p. a001107, 2010. [Online]. Available: http://cshperspectives.cshlp.org/content/2/2/a001107.full

[93] H. Wang, M. Karikomi, S. Naidu, R. Rajmohan, E. Caserta, H.-Z. Chen, M. Rawahneh, J. Moffitt, J. a. Stephens, S. a. Fernandez, M. Weinstein, D. Wang, W. Sadee, K. La Perle, P. Stromberg, T. J. Rosol, C. Eng, M. C. Ostrowski, and G. Leone, "Allele-specific tumor spectrum in pten knockin mice." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 11, pp. 5142–5147, 2010.

[94] A. Papa, L. Wan, M. Bonora, L. Salmena, M. S. Song, R. M. Hobbs, A. Lunardi, K. Webster, C. Ng, R. H. Newton, N. Knoblauch, J. Guarnerio, K. Ito, L. a. Turka, A. H. Beck, P. Pinton, R. T. Bronson, W. Wei, and P. P. Pandolfi, "Cancer-associated PTEN mutants act in a dominant-negative manner to suppress PTEN protein function," *Cell*, vol. 157, no. 3, pp. 595–610, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2014.03.027

[95] N. R. Leslie and J. Den Hertog, "Mutant PTEN in cancer: Worse than nothing," *Cell*, vol. 157, no. 3, pp. 527–529, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2014.04.008

[96] T. N. Turner, C. Douville, D. Kim, P. D. Stenson, D. N. Cooper, A. Chakravarti, and R. Karchin, "Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense distribution patterns," *Human molecular genetics*, vol. Accepted, 2015.

[97] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith, "Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms," *Journal of Mathematical Modelling and Algorithms*, vol. 5, no. 4, pp. 475–504, 2006. [Online]. Available: http://link.springer.com/10.1007/s10852-005-9022-1

[98] T. Sjöblom, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. V. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, and K. E. Bachman, "The Consensus Coding Sequences of Human Breast and Colorectal Cancers," *Science*, vol. 314, no. October, pp. 268–274, 2006.

[99] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik,

E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes." *Nature*, vol. 446, no. 7132, pp. 153–158, 2007.

[100] H. Carter and S. Chen, "Cancer-Specific High-Throughput Annotation of Somatic Mutations : Computational Prediction of Driver Missense Mutations Cancer-Specific High-Throughput Annotation of Somatic Mutations : Computational Prediction of Driver Missense Mutations," *Cancer Research*, vol. 69, no. 16, pp. 6660–6668, 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2763410&tool=pmcentrez&rendertype=abstract

[101] W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, and R. Karchin, "CHASM and SNVBox: Toolkit for detecting biologically important single nucleotide mutations in cancer," *Bioinformatics*, vol. 27, no. 15, pp. 2147–

2148, 2011.

[102] E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H.-R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp, "An overview of Ensembl," *Genome Research*, vol. 14, no. 5, pp. 925–928, 2004.

[103] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, 2007. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl842

[104] K. D. Pruitt, J. Harrow, R. a. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. DiCuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L.

Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. a. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes," *Genome Research*, vol. 19, no. 7, pp. 1316–1323, 2009.

[105] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. a. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–50, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16199517

[106] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," pp. 289 – 300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[107] N. Niknafs, D. Kim, R. Kim, M. Diekhans, M. Ryan, P. D. Stenson, D. N. Cooper, and R. Karchin, "MuPIT interactive: Webserver for mapping variant positions to annotated, interactive 3D structures," *Human Genetics*, vol. 132, pp. 1235–1243, 2013.

[108] "Jmol: an open-source Java viewer for chemical structures in 3D," http://www.jmol.org/, accessed: 2015-9-30.

[109] M. Ryan, M. Diekhans, S. Lien, Y. Liu, and R. Karchin, "LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures," *Bioinformatics*, vol. 25, no. 11, pp. 1431–1432, 2009. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp242

[110] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank." *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[111] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr981

[112] E. M. Gertz, Y.-K. Yu, R. Agarwala, A. a. Schäffer, and S. F. Altschul, "Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST." *BMC biology*, vol. 4, no. 1, p. 41, Jan. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1779365&tool=pmcentrez&rendertype=abstract

[113] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and

associated tools," *Briefings in Bioinformatics*, p. bbs038, Aug. 2012. [Online]. Available: http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbs038

[114] J. McGrath, S. Saha, D. Chant, and J. Welham, "Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality," *Epidemiologic Reviews*, vol. 30, no. 1, pp. 67–76, 2008. [Online]. Available: http://epirev.oxfordjournals.org/cgi/doi/10.1093/epirev/mxn001

[115] P. F. Sullivan, K. S. Kendler, and M. C. Neale, "Schizophrenia as a Complex Trait," *Archives of general psychiatry*, vol. 60, pp. 1187–1192, 2003.

[116] P. Lichtenstein, B. H. Yip, C. Bjork, Y. Pawitan, T. D. Cannon, P. F. Sullivan, and C. M. Hultman, "Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study." *Lancet*, vol. 373, no. 9659, pp. 234–239, 2009. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(09)60072-6

[117] J. van Dongen and D. I. Boomsma, "The evolutionary paradox and the missing heritability of schizophrenia," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 162, no. 2, pp. 122–136, 2013. [Online]. Available: http://doi.wiley.com/10.1002/ajmg.b.32135

[118] S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, S. M. Purcell Leader, D. M. Ruderfer, A. McQuillin, D. W. Morris, C. T. O'Dushlaine, A. Corvin, P. a. Holmans, S. Macgregor,

H. Gurling, D. H. R. Blackwood, N. J. Craddock, M. Gill, C. M. Hultman, G. K. Kirov, P. Lichtenstein, W. J. Muir, M. J. Owen, C. N. Pato, E. M. Scolnick, D. St Clair, P. Sklar Leader, N. M. Williams, L. Georgieva, I. Nikolov, N. Norton, H. Williams, D. Toncheva, V. Milanova, E. F. Thelander, P. Sullivan, E. Kenny, E. M. Quinn, K. Choudhury, S. Datta, J. Pimm, S. Thirumalai, V. Puri, R. Krasucki, J. Lawrence, D. Quested, N. Bass, C. Crombie, G. Fraser, S. Leh Kuan, N. Walker, K. a. McGhee, B. Pickard, P. Malloy, A. W. Maclean, M. Van Beck, M. T. Pato, H. Medeiros, F. Middleton, C. Carvalho, C. Morley, A. Fanous, D. Conti, J. a. Knowles, C. Paz Ferreira, A. Macedo, M. Helena Azevedo, A. N. Kirby, M. a. R. Ferreira, M. J. Daly, K. Chambert, F. Kuruvilla, S. B. Gabriel, K. Ardlie, and J. L. Moran, "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." *Nature*, vol. 10, no. AuGuST, pp. 8192–8192, 2009.

[119] H. Stefansson, R. a. Ophoff, S. Steinberg, O. a. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. H. Pietiläinen, O. Mors, P. B. Mortensen, E. Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Bø rglum, A. Hartmann, A. Fink-Jensen, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, Y. Böttcher, J. Olesen, R. Breuer, H.-J. Möller, I. Giegling, H. B. Rasmussen, S. Timm, M. Mattheisen, I. Bitter, J. M. Réthelyi, B. B. Magnusdottir, T. Sigmundsson, P. Olason, G. Masson, J. R.

Gulcher, M. Haraldsson, R. Fossdal, T. E. Thorgeirsson, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, B. Franke, E. Strengman, L. a. Kiemeney, Group, I. Melle, S. Djurovic, L. Abramova, V. Kaleda, J. Sanjuan, R. de Frutos, E. Bramon, E. Vassos, G. Fraser, U. Ettinger, M. Picchioni, N. Walker, T. Toulopoulou, A. C. Need, D. Ge, J. Lim Yoon, K. V. Shianna, N. B. Freimer, R. M. Cantor, R. Murray, A. Kong, V. Golimbet, A. Carracedo, C. Arango, J. Costas, E. G. Jönsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nöthen, M. Rietschel, P. M. Matthews, P. Muglia, L. Peltonen, D. St Clair, D. B. Goldstein, K. Stefansson, D. a. Collier, R. S. Kahn, D. H. Linszen, J. van Os, D. Wiersma, R. Bruggeman, W. Cahn, L. de Haan, L. Krabbendam, and I. Myin-Germeys, "Common variants conferring risk of schizophrenia," *Nature*, vol. 460, no. August, pp. 3–7, 2009. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature08186

[120] S. Ripke, A. R. Sanders, K. S. Kendler, D. F. Levinson, P. Sklar, P. a. Holmans, D.-Y. Lin, J. Duan, R. a. Ophoff, O. a. Andreassen, E. Scolnick, S. Cichon, D. St. Clair, A. Corvin, H. Gurling, T. Werge, D. Rujescu, D. H. R. Blackwood, C. N. Pato, A. K. Malhotra, S. Purcell, F. Dudbridge, B. M. Neale, L. Rossin, P. M. Visscher, D. Posthuma, D. M. Ruderfer, A. Fanous, H. Stefansson, S. Steinberg, B. J. Mowry, V. Golimbet, M. De Hert, E. G. Jönsson, I. Bitter, O. P. H. Pietiläinen, D. a. Collier, S. Tosato, I. Agartz, M. Albus, M. Alexander, R. L. Amdur, F. Amin, N. Bass, S. E. Bergen,

BIBLIOGRAPHY

D. W. Black, A. D. Bø rglum, M. a. Brown, R. Bruggeman, N. G. Buccola, W. F. Byerley, W. Cahn, R. M. Cantor, V. J. Carr, S. V. Catts, K. Choudhury, C. R. Cloninger, P. Cormican, N. Craddock, P. a. Danoy, S. Datta, L. de Haan, D. Demontis, D. Dikeos, S. Djurovic, P. Donnelly, G. Donohoe, L. Duong, S. Dwyer, A. Fink-Jensen, R. Freedman, N. B. Freimer, M. Friedl, L. Georgieva, I. Giegling, M. Gill, B. Glenthø j, S. Godard, M. Hamshere, M. Hansen, T. Hansen, A. M. Hartmann, F. a. Henskens, D. M. Hougaard, C. M. Hultman, A. Ingason, A. V. Jablensky, K. D. Jakobsen, M. Jay, G. Jürgens, R. S. Kahn, M. C. Keller, G. Kenis, E. Kenny, Y. Kim, G. K. Kirov, H. Konnerth, B. Konte, L. Krabbendam, R. Krasucki, V. K. Lasseter, C. Laurent, J. Lawrence, T. Lencz, F. B. Lerer, K.-Y. Liang, P. Lichtenstein, J. a. Lieberman, D. H. Linszen, J. Lönnqvist, C. M. Loughland, A. W. Maclean, B. S. Maher, W. Maier, J. Mallet, P. Malloy, M. Mattheisen, M. Mattingsdal, K. a. McGhee, J. J. McGrath, A. McIntosh, D. E. McLean, A. McQuillin, I. Melle, P. T. Michie, V. Milanova, D. W. Morris, O. Mors, P. B. Mortensen, V. Moskvina, P. Muglia, I. Myin-Germeys, D. a. Nertney, G. Nestadt, J. Nielsen, I. Nikolov, M. Nordentoft, N. Norton, M. M. Nöthen, C. T. O'Dushlaine, A. Olincy, L. Olsen, F. A. O'Neill, T. F. Ø rntoft, M. J. Owen, C. Pantelis, G. Papadimitriou, M. T. Pato, L. Peltonen, H. Petursson, B. Pickard, J. Pimm, A. E. Pulver, V. Puri, D. Quested, E. M. Quinn, H. B. Rasmussen, J. M. Réthelyi, R. Ribble, M. Rietschel, B. P. Riley, M. Ruggeri,

U. Schall, T. G. Schulze, S. G. Schwab, R. J. Scott, J. Shi, E. Sigurdsson, J. M. Silverman, C. C. a. Spencer, K. Stefansson, A. Strange, E. Strengman, T. S. Stroup, J. Suvisaari, L. Terenius, S. Thirumalai, J. H. Thygesen, S. Timm, D. Toncheva, E. van den Oord, J. van Os, R. van Winkel, J. Veldink, D. Walsh, A. G. Wang, D. Wiersma, D. B. Wildenauer, H. J. Williams, N. M. Williams, B. Wormley, S. Zammit, P. F. Sullivan, M. C. O'Donovan, M. J. Daly, and P. V. Gejman, "Genome-wide association study identifies five new schizophrenia loci," *Nature Genetics*, vol. 43, no. 10, pp. 969–976, 2011. [Online]. Available: http://www.nature.com/doifinder/10.1038/ng.940

[121] S. Ripke, B. M. Neale, A. Corvin, J. T. R. Walters, K.-H. Farh, P. a. Holmans, P. Lee, B. Bulik-Sullivan, D. a. Collier, H. Huang, T. H. Pers, I. Agartz, E. Agerbo, M. Albus, M. Alexander, F. Amin, S. a. Bacanu, M. Begemann, R. a. Belliveau Jr, J. Bene, S. E. Bergen, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, W. Byerley, W. Cahn, G. Cai, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, R. C. K. Chan, R. Y. L. Chen, E. Y. H. Chen, W. Cheng, E. F. C. Cheung, S. Ann Chong, C. Robert Cloninger, D. Cohen, N. Cohen, P. Cormican, N. Craddock, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, D. Demontis, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, N. Durmishi, P. Eichhammer, J. Eriksson, V. Escott-Price, L. Essioux,

A. H. Fanous, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, M. Friedl, J. I. Friedman, M. Fromer, G. Genovese, L. Georgieva, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, V. Golimbet, S. Gopal, J. Gratten, L. de Haan, C. Hammer, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. a. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, M. V. Hollegaard, D. M. Hougaard, M. Ikeda, I. Joa, A. Julià, R. S. Kahn, L. Kalaydjieva, S. Karachanak-Yankova, J. Karjalainen, D. Kavanagh, M. C. Keller, J. L. Kennedy, A. Khrunin, Y. Kim, J. Klovins, J. a. Knowles, B. Konte, V. Kucinskas, Z. Ausrele Kucinskiene, H. Kuzelova-Ptackova, A. K. Kähler, C. Laurent, J. Lee Chee Keong, S. Hong Lee, S. E. Legge, B. Lerer, M. Li, T. Li, K.-Y. Liang, J. Lieberman, S. Limborska, C. M. Loughland, J. Lubinski, J. Lönnqvist, M. Macek Jr, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, S. Marsal, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, O. Mors, K. C. Murphy, R. M. Murray, I. Myin-Germeys, B. Müller-Myhsok, M. Nelis, I. Nenadic, D. a. Nertney, G. Nestadt, K. K. Nicodemus, L. Nikitina-Zake, L. Nisenbaum, A. Nordin, E. OCallaghan, C. ODushlaine, F. A. ONeill, S.-Y. Oh, A. Olincy, L. Olsen, J. Van Os, P. Endophenotypes International Consortium, C. Pantelis, G. N. Papadimitriou, S. Papiol,

E. Parkhomenko, M. T. Pato, T. Paunio, M. Pejovic-Milovancevic, D. O. Perkins, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. a. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, U. Schall, C. R. Schubert, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, E. Sigurdsson, T. Silagadze, J. M. Silverman, K. Sim, P. Slominsky, J. W. Smoller, H.-C. So, C. a. Spencer, E. a. Stahl, H. Stefansson, S. Steinberg, E. Stogmann, R. E. Straub, E. Strengman, J. Strohmaier, T. Scott Stroup, M. Subramaniam, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, E. Söderman, S. Thirumalai, D. Toncheva, S. Tosato, J. Veijola, J. Waddington, D. Walsh, D. Wang, Q. Wang, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, E. H. M. Wong, B. K. Wormley, H. Simon Xi, C. C. Zai, X. Zheng, F. Zimprich, N. R. Wray, K. Stefansson, P. M. Visscher, W. Trust Case-Control Consortium, R. Adolfsson, O. a. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, A. D. Bø rglum, S. Cichon, A. Darvasi, E. Domenici, H. Ehrenreich, T. o. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, S. a. McCarroll, A. McQuillin, J. L. Moran, P. B. Mortensen, B. J. Mowry, M. M. Nöthen, R. a. Ophoff,

M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. C. Sham, P. Sklar, D. St Clair, D. R. Weinberger, J. R. Wendland, T. Werge, M. J. Daly, P. F. Sullivan, and M. C. ODonovan, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, pp. 421–427, 2014. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature13595

[122] S. H. Lee, T. R. DeCandia, S. Ripke, J. Yang, P. F. Sullivan, M. E. Goddard, M. C. Keller, P. M. Visscher, and N. R. Wray, "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs," *Nature Genetics*, vol. 44, no. 3, pp. 247–250, 2012.

[123] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kähler, L. Duncan, E. Stahl, G. Genovese, E. Fernández, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. E. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. N. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. a. McCarroll, and P. Sklar, "A polygenic burden of rare disruptive mutations in schizophrenia." *Nature*, vol. 506, no. 7487, pp. 185–90, Feb. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24463508

[124] M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer,

P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. a. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. a. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, and M. C. O'Donovan, "De novo mutations in schizophrenia implicate synaptic networks." *Nature*, vol. 506, no. 7487, pp. 179–84, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24463507

[125] A. C. Need, J. P. McEvoy, M. Gennarelli, E. L. Heinzen, D. Ge, J. M. Maia, K. V. Shianna, M. He, E. T. Cirulli, C. E. Gumbs, Q. Zhao, C. R. Campbell, L. Hong, P. Rosenquist, A. Putkonen, T. Hallikainen, E. Repo-Tiihonen, J. Tiihonen, D. L. Levy, H. Y. Meltzer, and D. B. Goldstein, "Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia." *American journal of human genetics*, vol. 91, no. 2, pp. 303–12, Aug. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3415532&tool=pmcentrez&rendertype=abstract

[126] B. Xu, I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. a. Gogos, and M. Karayiorgou, "De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia,"

Nature Genetics, vol. 44, no. 12, pp. 1365–1369, 2012. [Online]. Available: http://dx.doi.org/10.1038/ng.2446

[127] J.-H. Park, S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs, S. J. Chanock, and N. Chatterjee, "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries," *Nature Genetics*, vol. 42, no. 7, pp. 570–575, 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/ng.610

[128] G. M. Sullivan and R. Feinn, "Using Effect Size - or Why the P Value Is Not Enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–82, 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3444174&tool=pmcentrez&rendertype=abstract

[129] K. V. Morris and J. S. Mattick, "The rise of regulatory RNA," *Nature Reviews Genetics*, no. April, Apr. 2014. [Online]. Available: http://www.nature.com/doifinder/10.1038/nrg3722

[130] I. a. Qureshi and M. F. Mehler, "Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease," *Nature Reviews Neuroscience*, vol. 13, no. 8, pp. 528–541, 2012. [Online]. Available: http://dx.doi.org/10.1038/nrn3234

[131] G. Barry, "Integrating the roles of long and small non-coding RNA in brain

function and disease." *Molecular psychiatry*, vol. 19, no. 4, pp. 410–6, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24468823

[132] G. Barry, J. a. Briggs, D. P. Vanichkina, E. M. Poth, N. J. Beveridge, V. S. Ratnu, S. P. Nayler, K. Nones, J. Hu, T. W. Bredy, S. Nakagawa, F. Rigo, R. J. Taft, M. J. Cairns, S. Blackshaw, E. J. Wolvetang, and J. S. Mattick, "The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing," *Molecular Psychiatry*, vol. 19, no. 4, pp. 486–494, 2014. [Online]. Available: http://www.nature.com/doifinder/10.1038/mp.2013.45

[133] M. Chorev and L. Carmel, "The Function of Introns," *Frontiers in Genetics*, vol. 3, no. April, pp. 1–15, 2012. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fgene.2012.00055/abstract

[134] J. J. An, K. Gharami, G.-Y. Liao, N. H. Woo, A. G. Lau, F. Vanevski, E. R. Torre, K. R. Jones, Y. Feng, B. Lu, and B. Xu, "Distinct Role of Long 3 UTR BDNF mRNA in Spine Morphology and Synaptic Plasticity in Hippocampal Neurons," *Cell*, vol. 134, no. 1, pp. 175–187, 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0092867408007083

[135] U. Braunschweig, N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia, and B. J. Blencowe, "Widespread intron retention in mammals functionally tunes

transcriptomes." *Genome research*, vol. 24, no. 11, pp. 1774–86, 2014. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4216919&tool=pmcentrez&rendertype=abstract

[136] C. R. Sibley, W. Emmett, L. Blazquez, A. Faro, N. Haberman, M. Briese, D. Trabzuni, M. Ryten, M. E. Weale, J. Hardy, M. Modic, T. Curk, S. W. Wilson, V. Plagnol, and J. Ule, "Recursive splicing in long vertebrate genes." *Nature*, vol. 521, no. 7552, pp. 371–5, 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14466$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/25970246$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4471124

[137] M. Zylka, J. Simon, and B. Philpot, "Gene Length Matters in Neurons," *Neuron*, vol. 86, no. 2, pp. 353–355, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0896627315002834

[138] H. W. Gabel, B. Kinde, H. Stroud, C. S. Gilbert, D. a. Harmin, N. R. Kastan, M. Hemberg, D. H. Ebert, and M. E. Greenberg, "Disruption of DNA-methylation-dependent long gene repression in Rett syndrome," *Nature*, vol. 522, no. 7554, pp. 89–93, 2015. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14319

[139] M. Polymenidou, C. Lagier-Tourenne, K. R. Hutt, S. C. Huelga, J. Moran, T. Y. Liang, S.-C. Ling, E. Sun, E. Wancewicz, C. Mazur,

H. Kordasiewicz, Y. Sedaghat, J. P. Donohue, L. Shiue, C. F. Bennett, G. W. Yeo, and D. W. Cleveland, "Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43," *Nature Neuroscience*, vol. 14, no. 4, pp. 459–468, 2011. [Online]. Available: http://www.nature.com/doifinder/10.1038/nn.2779

[140] I. a. Swinburne, D. G. Miguez, D. Landgraf, and P. a. Silver, "Intron length increases oscillatory periods of gene expression in animal cells." *Genes & development*, vol. 22, no. 17, pp. 2342–6, 2008. [Online]. Available: http://genesdev.cshlp.org/content/22/17/2342.short

[141] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, a. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, and B. J. Frey, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, vol. 347, no. 6218, p. 1254806, 2014.

[142] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, "The RIN: an RNA integrity number for assigning integrity values to RNA measurements." *BMC molecular biology*, vol. 7, no. 1, p. 3, 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1413964&tool=pmcentrez&rendertype=abstract

[143] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[144] M. Gonen and E. Alpaydin, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011. [Online]. Available: ⟨GotoISI⟩://WOS:000293757900004

[145] B. Li and S. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, pp. 311–321, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0002929708004084

[146] B. E. Madsen and S. R. Browning, "A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic," *PLoS Genetics*, vol. 5, no. 2, p. e1000384, 2009. [Online]. Available: http://dx.plos.org/10.1371/journal.pgen.1000384

[147] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test." *American journal of human genetics*, vol. 89, no. 1, pp. 82–93, Jul. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3135811&tool=pmcentrez&rendertype=abstract

[148] J. A. Morris, G. Kandpal, L. Ma, and C. P. Austin, "DISC1 (Disrupted-In-Schizophrenia 1) is a centrosome-associated protein that

interacts with MAP1A, MIPT3, ATF4/5 and NUDEL: regulation and loss of interaction with mutation," *Human Molecular Genetics*, vol. 12, no. 13, pp. 1591–1608, 2003. [Online]. Available: http://www.hmg.oupjournals.org/cgi/doi/10.1093/hmg/ddg162

[149] L. M. Camargo, V. Collura, J.-C. Rain, K. Mizuguchi, H. Hermjakob, S. Kerrien, T. P. Bonnert, P. J. Whiting, and N. J. Brandon, "Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia," *Molecular Psychiatry*, vol. 12, no. 1, pp. 74–86, 2007. [Online]. Available: http://www.nature.com/doifinder/10.1038/sj.mp.4001880

[150] J. Costas, J. J. Suárez-Rama, N. Carrera, E. Paz, M. Páramo, S. Agra, J. Brenlla, R. Ramos-Ríos, and M. Arrojo, "Role of DISC1 Interacting Proteins in Schizophrenia Risk from Genome-Wide Analysis of Missense SNPs," *Annals of Human Genetics*, vol. 77, no. 6, pp. 504–512, 2013. [Online]. Available: http://doi.wiley.com/10.1111/ahg.12037

[151] E. K. Green, D. Grozeva, L. Forty, K. Gordon-Smith, E. Russell, a. Farmer, M. Hamshere, I. R. Jones, L. Jones, P. McGuffin, J. L. Moran, S. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock, "Association at SYNE1 in both bipolar disorder and recurrent major depression,"

Molecular Psychiatry, vol. 18, no. 5, pp. 614–617, 2013. [Online]. Available: http://www.nature.com/doifinder/10.1038/mp.2012.48

[152] F. O. Walker, "Huntington's disease." *Lancet*, vol. 369, no. 9557, pp. 218–28, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17240289

[153] J. Hui, L.-H. Hung, M. Heiner, S. Schreiner, N. Neumüller, G. Reither, S. a. Haas, and A. Bindereif, "Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing," *The EMBO Journal*, vol. 24, no. 11, pp. 1988–1998, 2005. [Online]. Available: http://emboj.embopress.org/cgi/doi/10.1038/sj.emboj.7600677

[154] F. Gebhardt, K. S. Zänker, K. S. Za, and B. Brandt, "Molecular Genetics : Modulation of Epidermal Growth Factor Receptor Gene Transcription by a Polymorphic Dinucleotide Repeat in Intron Modulation of Epidermal Growth Factor Receptor Gene Transcription by a Polymorphic Dinucleotide Repeat in Intron 1 *," *Journal of Biological Chemistry*, vol. 274, no. 19, pp. 13 176–13 180, 1999.

[155] G. Giannini, C. Rinaldi, E. Ristori, M. I. Ambrosini, F. Cerignoli, A. Viel, E. Bidoli, S. Berni, G. D'Amati, G. Scambia, L. Frati, I. Screpanti, and A. Gulino, "Mutations of an intronic repeat induce impaired MRE11 expression in primary human cancer with microsatellite instability,"

Oncogene, vol. 23, no. 15, pp. 2640–2647, 2004. [Online]. Available: http://www.nature.com/doifinder/10.1038/sj.onc.1207409

[156] C. Galindo, L. McIver, J. McCormick, M. Skinner, Y. Xie, R. Gelhausen, K. Ng, N. Kumar, and H. Garner, "Global Microsatellite Content Distinguishes Humans, Primates, Animals, and Plants," *Molecular Biology and Evolution*, vol. 26, no. 12, pp. 2809–2819, 2009. [Online]. Available: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msp192

[157] G. Highnam, C. Franck, a. Martin, C. Stephens, a. Puthige, and D. Mittelman, "Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles," *Nucleic Acids Research*, vol. 41, no. 1, pp. e32–e32, 2013. [Online]. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks981

[158] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich, "lobSTR : A short tandem repeat profiler for personal genomes lobSTR : A short tandem repeat profiler for personal genomes," *Genome research*, vol. 22, no. 6, pp. 1154–1162, 2012.

[159] O. Tange *et al.*, "Gnu parallel-the command-line power tool," *The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, 2011.

[160] J. Casbon, "PyVCF, a variant call format reader for Python," https://github.com/jamescasbon/PyVCF, accessed: 2015-10-05.

[161] M. Bayer, "SQLalchemy, the Python SQL toolkit and object relational mapper," www.sqlalchemy.org, accessed: 2015-10-05.

[162] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[163] N. Yannay-Cohen, I. Carmi-Levy, G. Kay, C. M. Yang, J. M. Han, D. M. Kemeny, S. Kim, H. Nechushtan, and E. Razin, "LysRS Serves as a Key Signaling Molecule in the Immune Response by Regulating Gene Expression," *Molecular Cell*, vol. 34, no. 5, pp. 603–611, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.molcel.2009.05.019

[164] S. G. Park, H. J. Kim, Y. H. Min, E.-C. Choi, Y. K. Shin, B.-J. Park, S. W. Lee, and S. Kim, "Human lysyl-tRNA synthetase is secreted to trigger proinflammatory response." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6356–6361, 2005.

[165] S. G. Park, P. Schimmel, and S. Kim, "Aminoacyl tRNA synthetases and their connections to disease," *Proceedings of the National Academy of Sciences*, vol. 105, no. 32, pp. 11 043–11 049, 2008.

[166] H. Javanbakht, R. Halwani, S. Cen, J. Saadatmand, K. Musier-Forsyth, H. Gottlinger, and L. Kleiman, "The Interaction between HIV-1 Gag and Human Lysyl-tRNA Synthetase during Viral Assembly," *Journal of*

Biological Chemistry, vol. 278, no. 30, pp. 27 644–27 651, 2003. [Online]. Available: http://www.jbc.org/cgi/doi/10.1074/jbc.M301840200

[167] J. Wenzel, J. L. Ouderkirk, M. Krendel, and R. Lang, "Class I myosin Myo1e regulates TLR4-triggered macrophage spreading, chemokine release, and antigen presentation via MHC class II," *European Journal of Immunology*, vol. 45, no. 1, pp. 225–237, 2015. [Online]. Available: http://doi.wiley.com/10.1002/eji.201444698

[168] C. Mele, P. Iatropoulos, R. Donadelli, A. Calabria, R. Maranta, P. Cassis, S. Buelli, S. Tomasoni, R. Piras, M. Krendel, S. Bettoni, M. Morigi, M. Delledonne, C. Pecoraro, I. Abbate, M. R. Capobianchi, F. Hildebrandt, E. Otto, F. Schaefer, F. Macciardi, F. Ozaltin, S. Emre, T. Ibsirlioglu, A. Benigni, G. Remuzzi, and M. Noris, "MYO1E Mutations and Childhood Familial Focal Segmental Glomerulosclerosis," *The New England Journal of Medicine*, vol. 365, no. 4, pp. 295–306, 2011.

[169] N.-S. Tzeng, Y.-H. Hsu, S.-Y. Ho, Y.-C. Kuo, H.-C. Lee, Y.-J. Yin, H.-a. Chen, W.-L. Chen, W. C.-C. Chu, and H.-L. Huang, "Is schizophrenia associated with an increased risk of chronic kidney disease? A nationwide matched-cohort study," *BMJ Open*, vol. 5, no. 1, pp. e006 777–e006 777, 2015. [Online]. Available: http://bmjopen.bmj.com/cgi/doi/10.1136/bmjopen-2014-006777

[170] E. Van Cauter, P. Linkowski, M. Kerkhofs, P. Hubain, M. L'Hermite-Balériaux, R. Leclercq, M. Brasseur, G. Copinschi, and J. Mendlewicz, "Circadian and sleep-related endocrine rhythms in schizophrenia." *Archives of general psychiatry*, vol. 48, no. 4, pp. 348–356, 1991.

[171] D. Pritchett, K. Wulff, P. L. Oliver, D. M. Bannerman, K. E. Davies, P. J. Harrison, S. N. Peirson, and R. G. Foster, "Evaluating the links between schizophrenia and sleep and circadian rhythm disruption," *Journal of Neural Transmission*, vol. 119, no. 10, pp. 1061–1075, 2012.

[172] K. Wulff, D. J. Dijk, B. Middleton, R. G. Foster, and E. M. Joyce, "Sleep and circadian rhythm disruption in schizophrenia," *British Journal of Psychiatry*, vol. 200, no. 4, pp. 308–316, 2012.

[173] N. Müller, J. K. Wagner, D. Krause, E. Weidinger, A. Wildenauer, M. Obermeier, S. Dehning, R. Gruber, and M. J. Schwarz, "Impaired monocyte activation in schizophrenia," *Psychiatry Research*, vol. 198, no. 3, pp. 341–346, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.psychres.2011.12.049

[174] J. Kowalski, P. Blada, K. Kucia, a. Madej, and Z. S. Herman, "Neuroleptics normalize increased release of interleukin- 1 beta and tumor necrosis factor-alpha from monocytes in schizophrenia." *Schizophrenia research*, vol. 50, no. 3, pp. 169–175, 2001.

[175] R. C. Drexhage, E. M. Knijff, R. C. Padmos, L. V. D. Heul-Nieuwenhuijzen, W. Beumer, M. a. Versnel, and H. a. Drexhage, "The mononuclear phagocyte system and its cytokine inflammatory networks in schizophrenia and bipolar disorder." *Expert review of neurotherapeutics*, vol. 10, no. 1, pp. 59–76, 2010.

[176] R. C. Drexhage, T. a. Hoogenboezem, D. Cohen, M. a. Versnel, W. a. Nolen, N. J. M. van Beveren, and H. a. Drexhage, "An activated set point of T-cell and monocyte inflammatory networks in recent-onset schizophrenia patients involves both pro- and anti-inflammatory forces." *The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum (CINP)*, vol. 14, no. 6, pp. 746–755, 2011.

[177] M. Keller, J. Mazuch, U. Abraham, G. D. Eom, E. D. Herzog, H.-D. Volk, A. Kramer, and B. Maier, "A circadian clock in macrophages controls inflammatory immune responses." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 50, pp. 21 407–21 412, 2009.

Vita

Dewey Kim, originally hailing from Portland, Oregon, received his B.A. in Physics at Reed College. Prior to enrolling into Johns Hopkins, Dewey joined the Vision Ergonomics Lab at Pacific Univeristy, learning Python and signal processing techniques while wrangling with EMG data. Once matriculated into the Johns Hopkins Biomedical Engineering Ph.D. program, Dewey caught the genomics bug, and is now currently attempting to apply data science techniques to discover new insights into cancer & psychiatric disease using next-generation sequencing data.

Beginning around the end of 2015, Dewey plans to join the Broad Institute in Cambridge, MA, where he will help discover new insights into the genetics of drug resistance in cancer.